From Disentangled Representation to Concept Ranking: Interpreting Deep Representations in Image Classification tasks*

Eric Ferreira dos Santos[1 [0000−0002−0408−5756]] and Alessandra Mileo[1 [0000−0002−6614−6462]]

Dublin City University, Collins Ave Ext - Whitehall, Dublin, Ireland
https://www.dcu.ie/courses/undergraduate/school-computing/computer-science
eric.ferreiradossantos2@mail.dcu.ie,alessandra.mileo@dcu.ie

Abstract. Deep Learning models such as Convolutional Neural Networks (CNNs) are particularly successful in computer vision tasks. They have proven to be tremendously effective and popular in the last decade, reaching great accuracy in tasks such as image classification and object recognition. Despite their success, it is well known that conveying what the model learnt to humans remains challenging. This is due to the fact that a CNN is still mostly a black-box model, and images are very rich input data. In this work, we build upon the idea of disentangled representation produced from a trained CNN, and explore how such disentangled representation can be used to describe what the model has learned in terms of semantic concepts. Specifically, we aim at providing a ranked list of the concepts that are related to both a specifica instance or image (local explainability) and a class (global explainability). In this preliminary work we use a simple linear classifier for concept ranking. Results are promising since we reached 95% precision at both local and global level. This indicates potential in developing our idea further by leveraging external knowledge bases to associate and validate specific properties and relations among the ranked concepts at both local and global level as discussed in the final section of this paper.

Keywords: Explainable AI · Disentangled Representation · Convolutional Neural Network

1 Introduction and Overview

Computer Vision is a branch of Artificial Intelligence that looks into how computer programs can interpret, represent, and act on visual inputs (such as pictures and videos). Deep Learning models such as Convolutional Neural Networks (CNNs) are specifically tailored to computer vision tasks, and in the last decade they have become remarkably successful and popular, achieving incredible accuracy in tasks such as image classification and object detection.

* Supported by Science Foundation Ireland - Grant No. 18/CRT/6223

Despite their success, CNNs are still mostly a black-box model, in that how and what the model learns is intelligible to the users and cannot be easily presented in human terms. The difficulty in generating a human-understandable explanation of the model outcome is hindering the use of CNNs in critical environments such as diagnostic imaging, disaster management and security surveillance to mention a few. In these scenarios it is crucial to understand how the model came to a given outcome and what the model has learned from the training data, not only to identify and correct mistakes, but also to detect potential bias in the data or the model.

The majority of approaches for interpreting directly the output of a trained CNN in a classification task have been focusing on the use of visual cues and more in general attention-based methods. For example, work in [8] and [5] have highlighted image pixels or areas contributing to a specific classification. [8] describes a technique for visualising how the model behaves in each layer for a particular image. Both approaches aid in localising which parts of the image were relevant to a specific class. However, because the image concepts are not declared in the image or the dataset, this visualisation does not represent them, and there is no guarantee that the model will highlight the same parts for other images in the same class.

In order to tackle the limitation of visual approaches, other explainability methods have been proposed in recent years, and the field of Explainable AI has began to be characterised in different survey papers.

Among others, [4] examines many strategies that employ textual justification when textual data is learnt and coupled with visual data, increasing the model's explainability. This method was utilised in the medical domain to clarify categorisation by combining image and textual diagnostics. Other techniques include simplification, which involves creating a white-box model from a complex model to achieve performance while simplifying the explanation. Feature relevance is another method which consists of considering each feature's value and using it to describe the learning process.

Another perspective in [2] is to explore using human expertise to explain how the model is learnt in a way that a layperson may comprehend, explained is rooted in real-world principles. This survey also provides links to the code for each approach discussed.

Based on the classification in these surveys, our approach would relate to the feature-relevant method, which in computer vision we would convert to real-world concepts. Ranking them, we intend to present the concepts more relevant from a singular image and to an entire class. In further work, we would combine this approach with the common-sense knowledge database, creating explanations that an AI system and a regular person can understand.

In this study, we will look at how a basic linear classifier can be used to rank concepts that characterise not only an instance, but more generally a class from local disentangled representations, which was not provided in previous works. We aim to provide a semantic explanation of what the model has learned in terms of the most relevant concepts. This is only the first step towards providing

an alternative human-understandable and self-explainable representation of a trained CNN model. In fact, we plan on building on the ability to not only identify semantic concepts as in disentangled representations, but also rank them based on their semantic relevance to an instance or a class (which is our key contribution in this work). Leveraging such ranking, we believe we can then go one step further in extracting semantic relations and subsequently learning logic rules from deep representations, as discussed in the final section of this paper.

The rest of the paper is structured as follows: in Section 2 we discuss related paper that specifically introduce and use the disentangled technique for improving model interpretability in image classification tasks; Section 3 describes our approach, specifically how we retrieved the local and global disentangled concepts from the trained model; our preliminary experimental evaluation and discussion of results is provided in Section 4, where we also outline how the evaluation should be extended and strengthened; we conclude in Section 5 presenting our ongoing research which builds upon the work in this paper towards a deeper understanding of the deep CNN model in a self-explainable and human-understandable way.

2 Related Work

Disentangled representation is a method that divides each characteristic (of an image) into carefully specified variables and encodes them as distinct dimensions. The idea is to emulate humans' fast intuitive process. [1] This method can characterise semantic concepts gained by a model throughout its training phase. This section will list the principal works that employed this strategy to characterise concepts learned in deep representations: Network Dissection, Decision Trees-based approach to learn disentangled filters and Concept Activation Vectors.

Network Dissection. Networks Dissection is a method for extracting meaning from each layer or filter, using the distillation approach to explain a CNN. Authors in [11] claim that a DNN may spontaneously learn disentangled representations. In order to demonstrate that, they developed a framework for connecting human notions to each filter in a CNN model (Figure 1). The objective is to provide meaningful labels to individual filters. The initial stage was to generate the Broden dataset, which contains pixel-annotated low-level notions like colours and high-level concepts like objects. They then used a trained model and passed through it the Broden dataset, to assess each filter and comparing the binary map from each picture and each filter activation map. If the convolutional filter is strongly activated in parts of the picture containing a human-labelled notion, authors claim that the filter is "searching for" that idea or concept.

[1] https://deepai.org/machine-learning-glossary-and-terms/disentangled-representation-learning

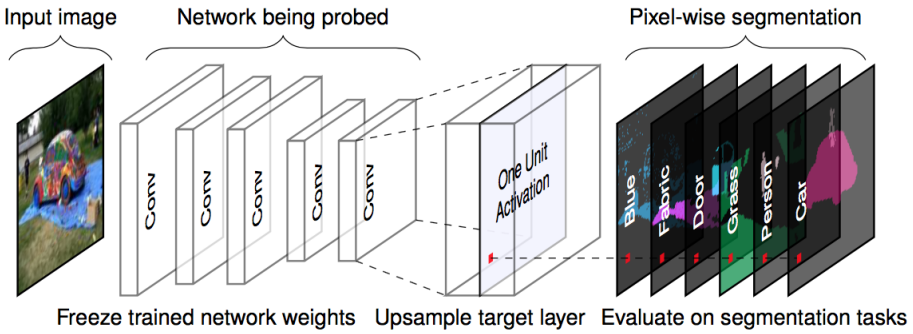

Fig. 1. Network Dissection framework ([11])

Examining different CNN designs, authors discovered certain important notions, such as the number of unique concepts for each layer in each architecture and the number of objects increasing into deeper convolutional layers.

Decision Trees Approach. In [10], the authors suggested learning a decision tree from a trained CNN, detailing the exact reasons for categorisation at a semantic level. The proposed technique describes which image components activate, which filters for categorisation and how much each part contributes. In this approach the authors use simplification method, to extract from a complex model a simple explanation.

The first part of the approach is training a CNN with disentangled filters on the high convolution layers to each filter learn a specific concept and associate each one to semantic meaning since they do not have any annotations of the concepts. This approach is presented in [9], where a loss function is applied for each filter in the top convolutional layer.

The trained disentangled filters extract information from each image and input it into a decision tree that understands its composition. There is no link between a filter and a human notion at this time, thus the authors use other datasets to assign a concept to a specific filter. They concentrated on a single topic (bird) and only used concepts relating to that issue. This method differs from [11] because it does not employ an extensive concept dataset to assign the concept to each filter. It can, however, be used to search for ideas that are not available in the Broaden dataset.

Concept Activation Vectors. Another relevant work in [3] proposes determining how human notions influence categorisation results. Authors defined and developed the CAV (Concept Activation Vectors) to transform a neural network's internal state into human-friendly notions. The method is useful because a human concept, such as "stripes," may be shown to impact the "zebra" class.

The core idea is retrieved from a trained model, a vector that characterises a particular concept, and then a directional derivative is used to assess concept sensitivity for a specific class. This method gives a local explanation for a specific

concept within a class, which may be required if the user already understands which concepts are applicable to a given class and wants to identify among a set of such concepts which ones are more descriptive for that class from the point of view of the deep representation, this validating which concepts among the given ones are affecting a classification most.

The papers and approaches discussed above provide some explanations for CNNs, incorporating human notions that might assist a non-specialist in determining how the model learnt a particular categorisation. These techniques, however, were not employed to describe a global classification, such as how the model understands a whole class. In this paper, we suggest using the disentangled approach described in [11] to determine how the classes may be interpreted using a global ranking of semantic concepts. We use a different dataset ([7]) to examine the ideas used to categorise a specific image. The model processed the dataset, and the total of each activation map of a unit in the final layer was used as input for an SVM classifier. We then use the top-ranked unit weighted by the SVM model for a particular image to order the identified concepts. Section 3 describes our naive technique in more detail.

3 Extracting Concepts for Action Classification

This section will outline the approach for concept extraction and ranking we propose in this paper.

The first step in this process is to extract semantic concepts about a class or a single instance from a trained CNN model using transfer-learning on an action classification task on a dataset containing forty actions. We decided to build upon previous research [11] that has already obtained promising results on semantic concept identification for a trained CNN model. One of the outcomes of such work is the ability to quantify how interpretable a CNN is by discovering how individual hidden units align to semantic concepts at each hidden layer. Concepts were identified as being part of six categories: object, part, material, colour, texture and scene. The architecture that identified more unique concepts among those tested was ResNet-152, as indicated in Figure 2; therefore this is the architecture we adopt.

We extend this technique for concept detection by identifying and connecting such concepts to output classes as well as individual input images (or instances). We chose to focus on the last CNN layer to maximise the number of unique high-level semantic concepts discovered; once more different concepts are harnessed, more concepts may be connected with local or individual examples.

We start from the semantic concepts identified by *Network Dissection*[11] from the trained CNN model and build upon the relationship between the convolutional filter and the semantic concept from the *Broden*[1] dataset. A transfer-learning approach makes it possible to adapt the *Network Dissection* method to be used on different data sets (and classification tasks) and determine which filters are activated by each new input. With this approach, the concepts learned from *Network Dissection* are extended to new input images: the top K highest-

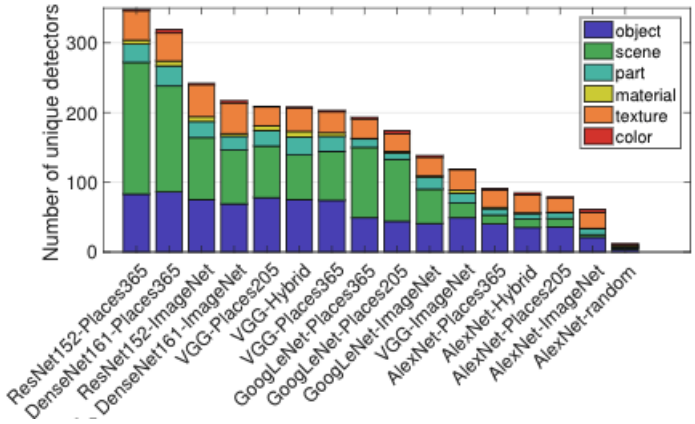

Fig. 2. Unique Detectors for each CNN Architecture ([11])

scoring filters for each input image are chosen as the identified concepts, considering the mean of each activation map.

Note that an activation map is a matrix that represents which image part was activated after the convolution function. It can be represented by a matrix $A_{M \times N}$ of the elements a_{ij},where $M = 1..i$ and $N = 1..j$. We define the mean of the activation map matrix $M_{Activation\_map}$ as follows:

$$M_{Activation\_map} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}}{\#E} \tag{1}$$

where $\#E$ is the number of elements of A.

We then rank the K filters from highest to lowest based on the mean activation map for each picture, assuming that the highest value identifies the most representative concept contained in an image. Once the model has identified other concepts for each image, the order of the pictures of the same class may be readjusted.

As a result, the approach produces as output a list of K different semantic concepts that are considered meaningful for each image. For the global concepts, a linear classifier with model-extracted features is applied to the same dataset for each class; subsequently, based on feature significance we determine which semantic concepts are relevant for the global separation. As mentioned previously, the dataset used for the investigation is the *Action40* dataset [7], which contains action photos labelled for 40 different action classes.

As a first metric for evaluation, we assess a simple precision from the ranked semantic concepts from local instances to the ranked semantic concepts for their class using the list of top K high-scored concepts from local and global examples. To do so, we compare the notions for each local example (image) belonging to a specific category to the top concepts that best linearly separate the class. We

consider the globally rated concepts to make sense with the local ones if at least one concept is offered between them. The formula for this can be expressed as:

$$P_c = \frac{\sum C_{l_c,g_c}}{\#L_c} \tag{2}$$

where P_c is the precision of the specific class c, $\sum C_{l_c,g_c}$ is the sum of the instances where the global and local shared at least one ranked semantic concept in the class c, and the $\#L_c$ is the number of local instances that belongs to the class c.

In this paper we assess the relationship between the top-ranked concepts from local and global examples using this metric. This gives us an indication of how well the global characteristics, separated linearly, reflect the semantic concepts acquired by the model for each class. We are aware this is a simplistic metric and we will discuss in the next section other possible variations that we will test and compare in future work.

The experimental evaluation of the extracted concepts following this approach will be presented in the next section.

4 Experimental Evaluation

As mentioned in section 3, we build upon the *Network Dissection* technique to extract local and global concepts. The task we consider in our investigation is action classification (*Action40* dataset)[2]. The concepts associated to each filter in the CNN model are provided by *Network Dissection*, which was trained on the *Imagenet* dataset[3], considering only a limited set of categories, namely *object*, *part*, *material* and *colour*. We only collected the concepts identified in the CNN's last layer, which created 162 distinct concepts. Then, using a transfer-learning approach, we used the *Action40* dataset to capture the concepts learnt for this data, based on the *Network Dissection* results.

The local semantic features from the new data were recovered using the mean of the activation map from each filter in the final layer as per Formula 1. The same formula was used for global concepts, i.e. concepts for each class, but this time it is used as feature extraction for the classification input. Following the intuition in [3] that meaningful higher-level concepts may be simpler to grasp, we used the SVM linear classifier to detect such concepts per class. We ran the algorithm using a 5-fold cross-validation, using the learning rate (C) equal to 0.001[4], and took the model with the best F1 score. The classification algorithm produced the confusion matrix in Figure 3, which displays the precision obtained for each class. The linear model achieved the precision of 80% in the class separation (classification task).

[2] http://vision.stanford.edu/Datasets/40actions.html
[3] https://www.image-net.org/download.php
[4] Best results from a grid-search technique using: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

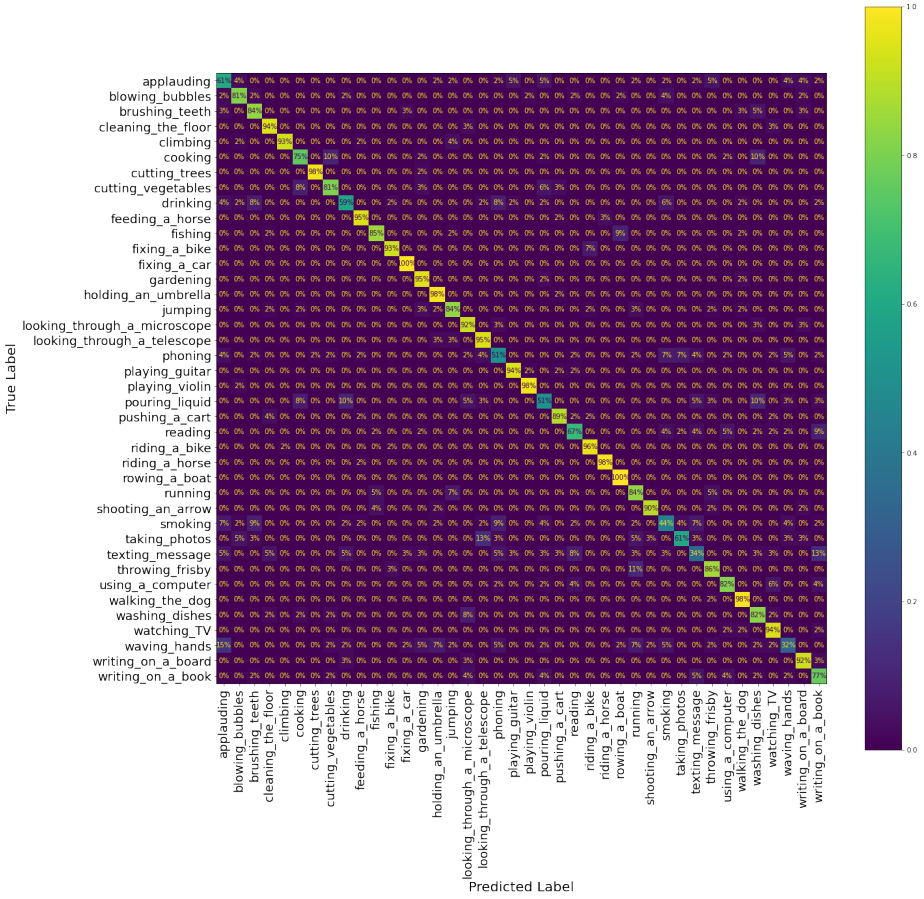

Fig. 3. Confusion Matrix from SVM classifier

Based on the top-ranked concepts extracted, we calculate the precision (Formula 2) between the images and their class in four different ways:

- Top 5 L - Top 1 G: The top 5 local concepts for each instance and the top 1 global concept for each class.
- Top 5 L - Top 5 G: The top 5 local concepts for each instance and the top 5 global concepts for each class.
- Top 5 L - Top 10 G: The top 5 local concepts for each instance and the top 10 global concepts for each class.
- Top 10 L - Top 10 G: The top 10 local concepts for each instance and the top 10 global concepts for each class.

The rationale behind varying the number of top concepts is to determine how many top global ranking concepts may best represent images from the same class. We discovered that a linear classifier will provide the feature relevance depending

on how successfully that feature separated the class, which is not always connected to the top local concepts, for example, from the same class. This means that in the model, the most common concepts provided in a group of images from the same class are not necessarily chosen as the best representative characteristics for that class. Therefore, we used this variation to precision between the local and global concepts belonging to the same class.

The first assessment evaluated whether the top one global concept from a particular class was present in the top five local concepts. Figure 4 shows that we could not identify a relevant precision for the majority of the classes utilising only the top 1 global concept. This behaviour supports the previous intuition by emphasising that the feature relevance in a linear classifier is aimed at the characteristics that best distinguish (or separate) the classes.

When we examine the precision between the top ten concepts in local and global instances (Figure 5), we can observe that the precision improves significantly, demonstrating that the global top ten concepts are represented in the local top ten concepts. Given that the model identified 162 different semantic concepts, and our technique could identify a mean precision of 95% between only ten ranking concepts, this is a significant result. The precision mean and standard deviation for all classes for each different number of global and local concepts are shown in Figure 6. Note that all the code is is available in an open repository on github [5] for reproducibility of results.

It is important to note that we only use precision as a quantitative measure for our concept ranking method. This is because in this instance we only check if global concepts are presented in local ones. Additional measures like recall and F1 would not change this result but we agree that they could provide other interesting insights. In order to further validate the proposed technique, we will not only explore the insights provided by using alternative evaluation metrics, but also compare results across different benchmark datasets.

In order to illustrate our outcome qualitatively with an example, we chose one of the greatest and lowest precision classes, “cutting_trees” and “phoning”, respectively. The class ”cutting_trees” had a significant separation result from the linear classifier (98%) and obtained 100% precision between the global and local concepts. Based on the feature significance from the linear model, the global concepts for this class are: “snow”, “tree”, “bird”, “motorbike”, “house”, “bicycle”, “plant” and “hand”. When we look at all of the photographs in the same class, the top ten local concepts are: “house”, “tree”, “plant”, “bird”, “person”, “bicycle”, “hand”, “motorbike”, “snow” and “food”.

This result demonstrates that there is an interesting overlap between global and local concepts for the class “cutting_trees,” which we can use to describe what the model learned as the pattern of this class. Simultaneously, we may manually check that the presented notions appear plausible when we consider the activity of cutting the tree and its images on the dataset. This is just an intuition, as we said, and a more systematic evaluation (either manually by humans or automatically via labels) should be conducted.

[5] https://github.com/EricFerreiraS/disentangled_representation-concept_ranking

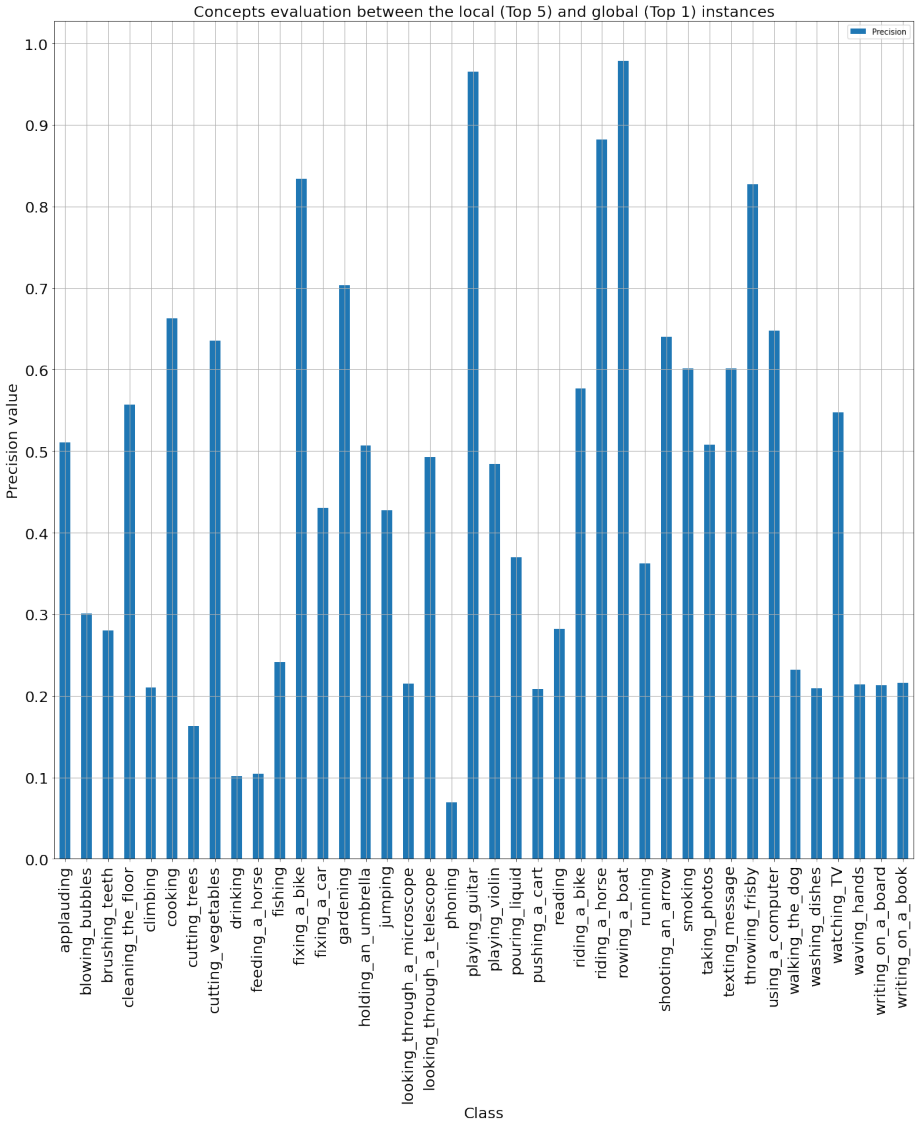

Fig. 4. Top 5 local concepts X Top 1 Global concept

When we look at the "phoning" class, the linear classifier did not produce an flattering result (precision of 51%), and when compared to the global and local concepts, the result was the poorest in the method (about 67%). This result might indicate two possibilities: the linear model did not segregate the concepts properly (an issue with the linear model) or there is a lack of concepts that could

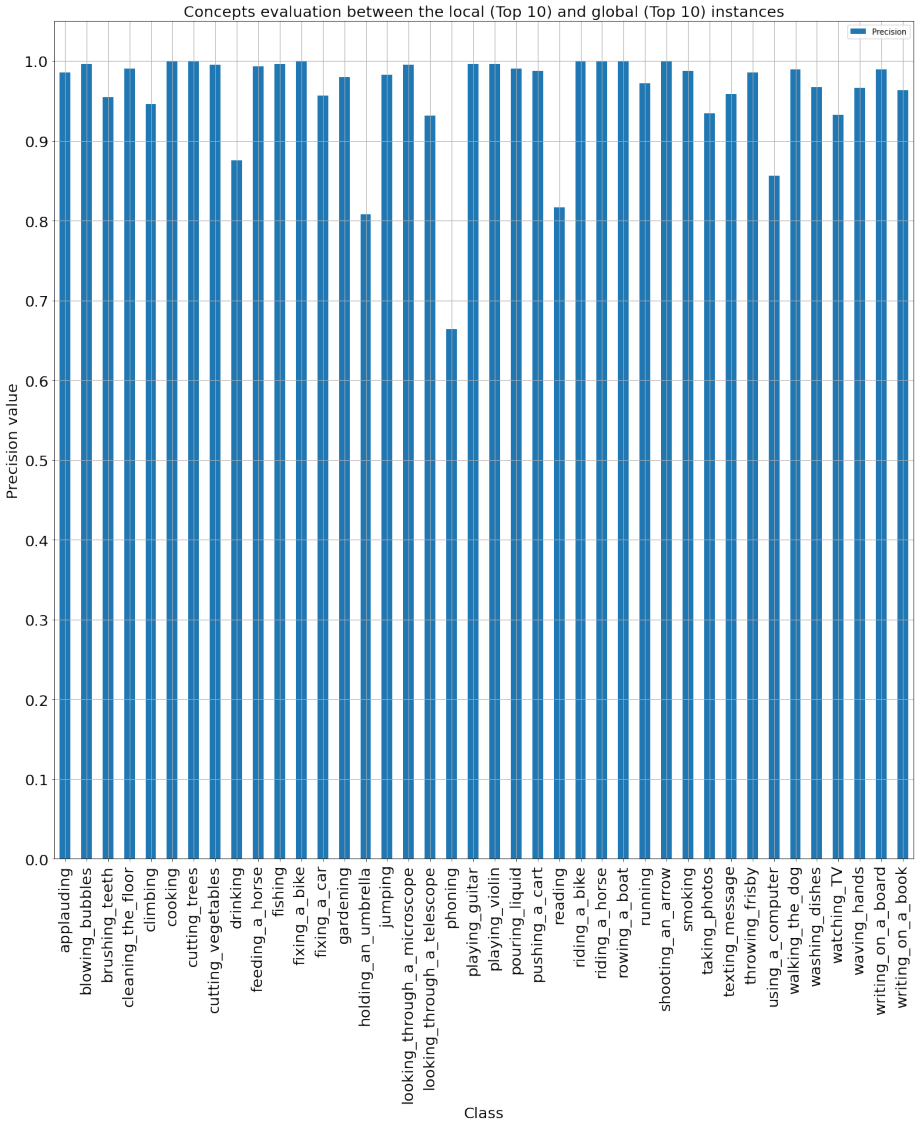

Fig. 5. Top 10 local concepts X Top 10 Global concept

better describe this class (issue with concept generation). These concerns will be investigated upon in future work.

To summarise, our quantitative experimental analysis so far showed that our approach was able to successfully retrieve the top 10 concepts from disentangled representation that best characterise the local instances (as per *Network Dissection*) as well as the global instances. We assessed the method by comparing the

existence of concepts in local and global occurrences. In the next section we will present the ongoing research we are conducting in this area and our next steps.

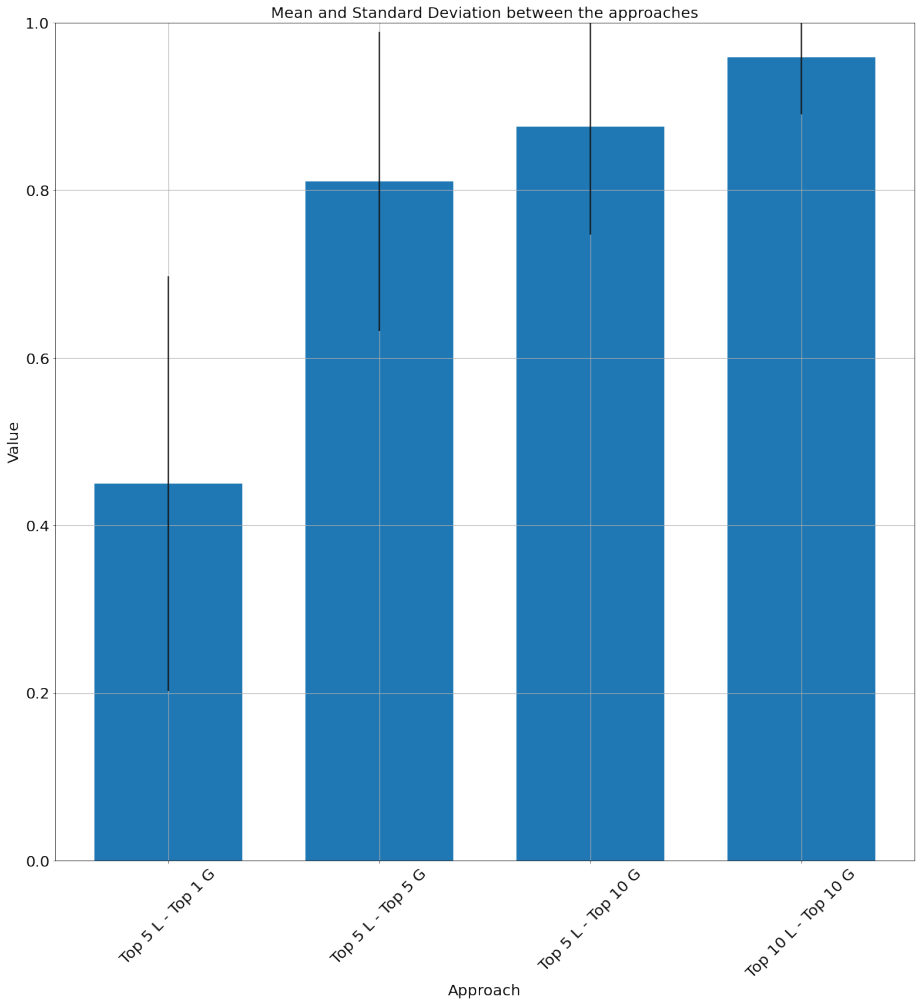

Fig. 6. Mean and Standard Deviation between the Precision

5 Ongoing Work

CNN has shown impressive accuracy in computer vision applications, but the absence of an explanation for what the model learnt remains an open challenge for its adoption in high-risk scenarios. This study investigated the potential of

building upon disentangled representations to provide a semantically meaningful interpretation of classification results produced by a CNN in terms of relevant semantic concepts. We demonstrate how even using a linear classifier such as SVM, we are able to meaningfully rank top ten concepts that characterise not only an instance, but more generally a class from local disentangled representations.

We define and test a method for extracting not only the top local concepts but also global ones. We demonstrated that we can identify the top concepts for an image of a given class, and that these are the same concepts necessary to best separate this class. For example, with a precision of 95% between the concepts presented in the images and their class, we have that images categorised as "riding a bike" contain the top local concepts "bicycle" and "wheel", and the same top concepts were necessary to separate this class according to the linear classification. As a result, we argue that the model has learned those concepts related to a specific class (and instances of that class). This paves the way for a concept-driven explanation of classification results using disentangled representations, although different challenges lie ahead.

For example, we notice that no semantic relationship between extracted concepts can be extracted with our method alone. To this aim, we believe leveraging an external knowledge base can aid in detecting semantic relationships between those concepts. We are currently investigating on the use of *Conceptnet*[6], a common-sense knowledge graph database, to acquire those relationships, thus improving the transparency and interpretability of what the model has learned.

Another key challenge is that the human expert's capacity to explain outcomes semantically is limited due to a lack of information regarding causal linkages between those concepts and their relationships. To tackle this we are also investigating the possibility of leveraging the extracted concepts and relationships to learn symbolic rules about causality, therefore offering a structural and human-like way of explaining the results of a decision made by the model.

References

1. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3319–3327 (2017). https://doi.org/10.1109/CVPR.2017.354
2. Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W.: Explainable ai methods-a brief overview. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. pp. 13–38. Springer (2022)
3. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)
4. Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review pp. 1–66 (2021)
5. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
6. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: An open multilingual graph of general knowledge. In: Thirty-first AAAI conference on artificial intelligence (2017)
7. Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L., Fei-Fei, L.: Human action recognition by learning bases of action attributes and parts. In: 2011 International conference on computer vision. pp. 1331–1338. IEEE (2011)
8. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
9. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8827–8836 (2018)
10. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting cnns via decision trees. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6261–6270 (2019)
11. Zhou, B., Bau, D., Oliva, A., Torralba, A.: Interpreting deep visual representations via network dissection. IEEE transactions on pattern analysis and machine intelligence **41**(9), 2131–2145 (2018)