Towards Understanding Deep Representations in CNN: from Concepts to Relations Extraction via Knowledge Graphs

Eric Ferreira dos Santos[1,*], Alessandra Mileo[1]

[1]Dublin City University, Collins Ave Ext - Whitehall, Dublin, Ireland

Abstract

The ability to understand deep representations from trained Convolutional Neural Networks (*CNN*) in image classification tasks is still limited when it comes to effectively justifying the reasons behind a given outcome. This is due to the fact that most approaches focus on low-level features (such as pixels to generate saliency maps) while human understanding is based on concepts and relations among those concepts. To address this problem, we propose an approach that aims to extract high-level human concepts and their relations from deep learning models by combining disentangled representations and Commonsense Knowledge Graphs and relying on textual descriptions of visual relations as ground truth for evaluation. The concept extraction phase leverages *Network Dissection* as a disentangled representation approach to collect global and local concepts learned by a trained *CNN* combined with a linear classifier. The relation extraction step uses a *CSKG* as commonsense knowledge graph to find relations between those concepts. The visual relation dataset *Visual Genome* is used as a ground truth to validate the known relations. Based on relations coverage between the local and global features, our approach paves the way to understand what a *CNN* learned in a way that can be easily interpreted by humans, presenting the importance of specific concepts and relations for a given classification task.

Keywords

Explainable AI, Computer Vision, Knowledge Graph, Convolutional Neural Network

1. Introduction and Motivation

Deep learning models, specifically Convolutional Neural Networks (*CNNs*) [1], have revolutionised computer vision applications such as image classification, object detection, and segmentation. While these models have attained cutting-edge performance on a variety of tasks, their intrinsic complexity and limited interpretability beyond visual cues pose substantial barriers to their use in real-world applications where trustworthiness is key.

The field of Explainable AI (*XAI*) has gained significant traction as researchers aim to develop reliable methods to unveil the deep neural network's decision process. In image classification, *XAI* techniques help identify which image features are critical for a model's outcome. Visual approaches, such as those described in [2, 3] focus on relating outputs with image features, but they require considerable human interpretation and can produce inconsistent results across different samples in the dataset. As a result, alternative strategies that go beyond visual explanations, including textual justifications and feature relevance, have emerged. These include approaches for combining textual and visual data, simplifying complex models, and utilising human-understandable concepts [4].

One of the key challenges in supporting human understanding of deep neural networks is the ability to effectively disentangle low level features from high level concepts. This has motivated approaches such as the one in [5], where disentangled concepts are ranked and considered as high-level features at both local and global level. In order to improve understanding of deep representation in image classification tasks beyond mere concept ranking, we want to enable discovery of relations among disentangled concepts. To this aim, this research paper proposes an approach that leverages external

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
*Corresponding author.
eric.ferreiradossantos2@mail.dcu.ie (E. F. d. Santos); alessandra.mileo@dcu.ie (A. Mileo)
0000-0002-0408-5756 (E. F. d. Santos); 0000-0002-6614-6462 (A. Mileo)

knowledge graphs as well as textual descriptions for discovery, ranking and validation of semantic relations among concepts top ranked concepts.

Our main contributions is a method that takes as input disentangled concepts learned from a *CNN* trained for an image classification task, and leverages a commonsense knowledge graph to extract meaningful relations among concept pairs. Our validation approach relies on the *Visual Genome* dataset[1], where concepts in images are related to their textual descriptions.

The remainder of the paper is organised as follows: Section 2 presents related works on concept extraction from *CNN* and the use of knowledge graph for modelling relations. Section 3 describes our approach; Section 4 discusses our experimental evaluation; and Section 5 concludes by discussing challenges and promising directions.

2. Related Work

When it comes to image classification as computer vision task, most published works so far focus on the use of *CNN* architectures to achieve the best performance. A more recent approach that is recently emerging, known as Vision Transformers (*ViT*) [6], relies on attention mechanisms and their success in natural language processing [7]. Although *ViT* outperforms *CNN* models in some cases, the difference is not significant considering the additional time and data needed for training and the risk of increased bias [8]. For these reasons, in this paper, we focus on using *CNNs* for our investigation.

When it comes to understanding the inner working of the model, among the approaches for explainability we focus on feature relevance [9]: specifically, we are interested in how to effectively translate relevant features into semantic concepts which humans can understand, and how to leverage knowledge graphs to elicit relations among such concepts.

A concept refers to a high-level, human-understandable abstraction, such as the colour "black" or the object "bicycle", which are defined by humans and do not require any additional information for a human to understand [10]. We can define a relation as a connection or association between two or more concepts. This connection can be based on factors such as causality, similarity, proximity, or functional dependence. For example, the concepts "bicycle" and "wheel" can be connected by a relation such as "a wheel is part of a bicycle", as well as "a bicycle contains a wheel.

Several approaches for *DNN*-based Relation Extraction have been proposed in Natural Language Processing [11]. However, the use of knowledge graph is still under-investigated as a promising direction to get direct access to the relational knowledge for explanation generation and reasoning, as well as obtain robust outcomes that are less dependent of the input data [12]. To the best of our knowledge, none of these approaches have been explored for image data specifically. The use of knowledge graphs to bridge the gap between visual concepts and semantic relations is not assuming the presence of textual labels or captions (although we need them for evaluation), and this characterises the new angle of our approach, but also determines the lack of a comparative baseline.

For this reason, as part of related work, we focus on approaches that tackle each step individually, namely: methods for concept extraction from deep models and the use of knowledge graphs to discover relations between semantic concepts (mostly in text).

2.1. Concept Extraction

When it comes to explainability, it is necessary to go beyond feature-relevance [13], which focuses on relating low-level features with high-level human concepts [14] assuming that each filter/neuron can independently be responsible for learning one or more concepts.

Network Dissection [15] is one of such approaches for disentangled representation, which assign relevant labels to each filter of a *CNN*. To achieve that, authors rely on the *Broden* dataset, composed of pixel-annotated low-level concepts such as colours and high-level concepts such as objects. A trained *CNN* uses the *Broden* dataset to compare the binary map from each image with each filter activation

[1] https://homes.cs.washington.edu/~ranjay/visualgenome/index.html

map: if the convolutional filter is substantially activated in areas of the image containing a given human-labelled concept, the filter is considered to be "looking for" that concept for a particular test image.

A different approach presented in [16] proposed a method to learn a decision tree which approximates the hierarchy of concepts learned by a trained *CNN*. Unlike *Network Dissection* [15], this approach does not rely on object-part annotations and is therefore only an approximation of the *CNN* decision process. In addition, it can be subject to noise when filters are activated by unrelated visual concepts.

The approach in [17] suggests investigating how specific human concepts impact classification outcomes based on Concept Activation Vectors (*CAV*). The general idea of *CAV* is to train a linear (binary) classifier to separate each specific concept and use a directional derivative to assess concept sensitivity for a particular class. This method identifies the relevance of a particular concept for the selection of a class, but it is required that the user already knows which concepts are more descriptive for that class to verify which ones affect a classification the most.

Despite these approaches rely on exposing a set of (pre-defined or dataset dependent) human concepts that can help identify what the model learned for a specific sample, they do not provide a global characterisation of how the model generalises to an entire class. An improvement in this direction is presented in [5], where authors combine the *Network Dissection* approach for disentangled representations with a linear classifier to understand which concepts the model learned both locally and globally. Local concepts are those related to each input image, while global concepts are generalised over an entire class. Specifically, authors in [5] use *Network Dissection* to identify the top-k local concepts for each image of a given class, and train a linear *SVM* classifier to collect the best features that were more important for the classification.

We believe human explanations are characterised by relations between concepts, as opposed to concepts in isolation. To this aim, in this paper we focus on going from concepts to relations, and we suggest doing that by leveraging approaches to relation extraction from knowledge graphs, applied to a ranked list of concepts such as those generated in [5]. Relevant work on relation extraction and knowledge graphs is presented in the next subsection.

2.2. Relation Extraction via Knowledge Graphs

A knowledge graph (*KG*) is a structured representation of knowledge containing entities, attributes, and relations. *KGs* capture the meaning of unstructured data and provide a semantic framework for intelligent systems [18]. *KGs* have been used in various applications, including question answering, information extraction, and entity recognition [19]. They have also been used to combine data from numerous sources, making complex data more accessible to reason about automatically. Machine learning and artificial intelligence advancements have fuelled the creation of *KGs* from large-scale data sources by using algorithms for knowledge extraction [20].

Commonsense *KGs* are specifically useful for reasoning about entities and their relations, which people consider intuitive, and this knowledge is critical for artificial intelligence applications because it has the potential to help machines understand and reason about the world like humans do.

The CommonSense Knowledge Graph (*CSKG*) [21] facilitates the use of such knowledge and it is a resource that uses seven very diverse and disjoint sources: a text-based commonsense knowledge graph *ConceptNet*[2], a general-purpose taxonomy *Wikidata*[3], an image description dataset *Visual Genome* [22], a procedural knowledge source *ATOMIC*[4], and three lexical sources: *WordNet*[5], *Roget*[6], and *FrameNet*[7]. Through the combination of different sources, the *CSKG* offers a variety of nodes (objects or concepts) and edges (relations) in order to provide a common sense knowledge base for reasoning.

[2]https://conceptnet.io/
[3]https://www.wikidata.org/wiki/Wikidata:Main_Page
[4]https://huggingface.co/datasets/allenai/atomic
[5]https://wordnet.princeton.edu/
[6]https://www.gutenberg.org/ebooks/22
[7]https://framenet.icsi.berkeley.edu/

KG have been employed to develop automatic methods for understanding the relations between objects in a picture. For example, *ConceptNet* was used in the images following the classification work in [23] to identify sample-specific relations. However, the method to align the concepts is mostly manual [24], and it does not explain how the model learned those concepts and relations. Authors in [25] also suggest an ontology-based approach to identify objects and their relations using a *KG*. However, the *KG* alignment with the concepts and how the deep representation learned their relations is still underinvestigated.

Concept-based explanation approaches discussed earlier in this section provide numerous ways for interpreting what the *CNN* model learned, but they do not provide any tool or approach for automatic relation extraction, which we aim to achieve using Knowledge Graphs.

This paper proposes a solution combining these two steps to go from concepts learned by a trained *CNN* to relations via knowledge graphs. Building upon approaches for concept extraction such as those in [15, 5] we aim to extract and validate relations among those concepts using *CSKG* as Commonsense Knowledge Graph. Our approach is detailed in the next section.

3. Overall Methodology for Concepts and Relation Extraction and Validation

The proposed approach has three phases, as illustrated in Figure 1: concept extraction, relation extraction and relation validation. We rely on *Network Dissection* as in [15, 5] for concept extraction, while *CSKG*[8] is used as knowledge graph to extract relations between concepts and the *Visual Genome* dataset is used to validate learned relations. It is important to emphasise that each phase of the approach can use different knowledge or dataset in order to extract the concepts and relations, as well as to validate those relations. In the remainder of this section we describe the datasets and knowledge graph used in our experiments, and we detail each of these phases separately.

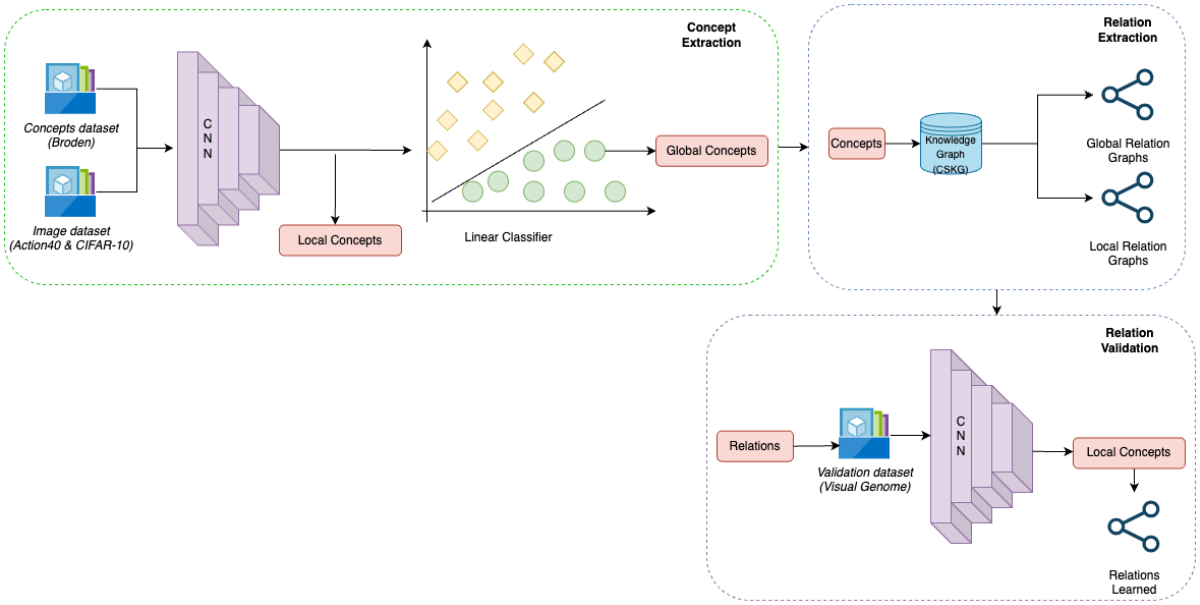

Figure 1: Conceptual diagram of the approach

[8] https://cskg.readthedocs.io/en/latest/

3.1. Datasets and Knowledge Graph

The *Broden* dataset [15] contains 63,305 images[9] with detailed annotations for objects, parts, materials, and scenarios, making it ideal for associating filters with high-level concepts in the concept extraction step. In our experiments, we replaced the last fully connected layer of a pre-trained model for action recognition based on the *Action40* dataset [26], which includes 9,532 images across 40 action classes; we also use the well-known *CIFAR-10* [27] dataset, containing 60,000 small images across ten classes, as a benchmark for our approach performance.

For Relation extraction, we rely on the *CSKG* (Commonsense Knowledge Graph) [21], which provided a framework for investigating common-sense reasoning by exploring the relations between concepts, with over 2 million concepts and 58 relations. The relations presented in this dataset are taxonomical and lexically based, combined from different sources. In addition, the *Visual Genome* [22] dataset, containing 101,174 images which include around 42,000 distinct relations, was used for interpreting visual scenes, with its extensive annotations of objects, relations, and scene attributes, aiding in validating relations learned by the model.

3.2. Concept Extraction

In the concept extraction phase, we apply *Network Dissection* on a *ResNet-152* model [28], trained on *ImageNet*[10] similarly to [5], and we use transfer learning to adapt the model weights to the *Action40* [26] and *CIFAR10* [27] datasets. We identify the most relevant concepts for each input image by taking the K highest-scoring filters in the last *CNN* layer, based on the mean of each activation map.

The linear SVM classifier is then used to identify the top K features per class (referred to as global concepts); we then evaluate which semantic concepts better separate classes based on feature significance. Authors in [5] had experimentally identified $K = 10$ as the value obtaining the best (95%) precision for both local and global concepts, therefore we used the same value for K.

3.3. Extracting Candidate Relations

Given the top K local and global concepts identified in the first phase, we query the *CSKG* knowledge graphs for candidate relations among all combinations of concept pairs in the top K. This step relies on query pattern matching via *KGTK*, a python library for *KG* manipulation[11]. Note that in *CSKG* nodes represent concepts and edges represent relations[12]. In this preliminary investigation, and in order to reduce the combinatorial complexity of this step, we only focused on direct relations between concepts: the algorithm takes a concept pair as input and returns as output only direct edges between the nodes in the concept pair. Note that at this point nodes are possible candidate relations based on commonsense knowledge, and therefore we still need to verify whether the deep representation has learned such relations. We do that as discussed in the next subsection.

3.4. Visual Validation of Candidate Relations

The *Visual Genome* dataset [22] contains pictures with tagged relations identified by bounding boxes. In order to identify which relation our model is likely to have learned, we proceed as follows:

1. for every candidate relations R_i among concepts C_1 and C_2 extracted from *CSKG*, we identify all images I_j in *Visual Genome* representing R_i[13];
2. every image I_j for a given relation R_i is passed through the model used for concept detection, with the hypothesis that if the two relevant concepts C_1 and C_2 for R_i are among the top ten activated filters, the network has likely learned the corresponding relations;

[9] http://netdissect.csail.mit.edu/data/broden1_227.zip
[10] https://www.image-net.org/
[11] https://kgtk.readthedocs.io/en/latest/transform/query/
[12] https://github.com/commonsense/conceptnet5/wiki/Relations
[13] Here we use string matching between *CSKG* and *Visual Genome* concepts

3. due to differences in relation names between *CSKG* and *Visual Genome*, we relaxed the exact string matching approach by comparing the results obtained with the use of Named-Entity Resolution (*NER*) approaches from Natural Language Processing (namely from *spacy.io*[14] and *NLTK*[15]) and the well known *word2vec*[16].

The reason we selected *Visual Genome* as a more robust approach to validation as opposed to Large Language Models such as *GPT-4*[17] or *Gemini*[18], is the reduced risk of hallucinations [29], as well as the ability to have access to a visual representation of relations that we could use as a ground-truth to relate concepts to disentangled filters.

4. Experimental Evaluation

Our experiments were conducted on a machine running Linux Mint 21.2, with 48 CPUs and 128 GB of RAM. We rely on two main Python libraries: Pytorch[19] for training and testing the models and *KGTK* to work with the knowledge graph. The code used in this research is available on the GitHub repository [20].

The concepts extraction phase on *ResNet-152* pre-trained on the *Imagenet* dataset resulted in 162 different concepts including object, part, material, and colour. We started from this pre-trained model and applied transfer learning, freezing all the trained layers and replacing the fully connected layers with the linear classifier. This enabled us to apply concept extraction, relation extraction and validation (with string matching and with Named-Entity Resolution or *NER*) on two different datasets, namely *Action40* and *CIFAR-10*.

The results of our analysis are presented in Table 1, where we can see the number of global (# R (Global)) and local (# R (Local)) unique relations extracted using *CSKG*, the percentage of the *Visual Genome* images containing the K local and global concepts extracted from each dataset (% VG Images), and the total number of unique relations learned using *Visual Genome* (#R (VG)). Note that the percentage of images from *Visual Genome*, and the relations learned are not distinguished as local and global, as they relate to the entire model when considering the overlap between local and global.

Table 1
Candidate local and global relations from *CSKG*, Percentage of the Visual Genome images that contain the same pair of concepts in local and global relations from *CSKG*, and Relations found using the *Visual Genome* only.

	# R (Global) CSKG	# R (Local) CSKG	% VG Images	# R (VG)
Action40	339	2,495	12%	3,434
CIFAR-10	137	2,176	14%	2,566

Based on the relations learned (Table 1, last column), we can now look for relations that are present in both *CSKG* and *VG*, both local and global. Table 2 shows these common relations found by simple string matching (#R) and by relaxing the matching using *NER* approaches (#R_NER). We consider the sum of unique relations extracted by any of the three *NER* approaches. We did not consider the contribution of each single approach for this analysis. Table 2 also presents the percentage of relations validated from *Visual Genome* more than once, that were also present in *CSKG* (%R_VAL) separated in global and local relations respectively. We use this threshold (relation validated more than once) to increase the likelyhood that the relation was actually learned. This means, for example, that starting from the

[14]https://spacy.io/models/en#en_core_web_lg
[15]https://www.nltk.org/index.html
[16]https://radimrehurek.com/gensim/models/word2vec.html
[17]https://openai.com/index/gpt-4/
[18]https://gemini.google.com/app
[19]https://pytorch.org/get-started/locally/
[20]https://github.com/EricFerreiraS/relation_extraction_AICS24

candidate global relations (Table 1 #R (Global)), extracted from the combination of the top K features and the *CSKG* in the candidate relation extraction phase, only 8% were validated as learned by the model for the *Action40* dataset.

Table 2
The number of relations found, locally and globally (R), their relaxation with NER (R_NER), and the percentage of global and local relations candidates validated.

	# R (Global)	# R_NER (Global)	# R (Local)	# R_NER (Local)	% R_VAL (Global)	% R_VAL (Local)
Action40	7	51	23	166	8%	9%
CIFAR-10	1	15	23	118	2%	3%

We observe that the there is a high number of local relations, but when validated across the instances of a class, only a few of those relations are likely to influence the classification task. Our method allows us to identify such global relations reducing noise.

In order to capture a more fine grained view of this phenomenon per class, we define the notion of coverage which measures how many local relations (for all images of a given class) are also present as global relations for that class. These global relations are relations among concepts specific to that class, as they best separate that class's feature space. If at least one relation is given, the globally rated relations make sense with the local ones. The coverage formula is as follows:

$$Coverage_c = \frac{\sum C_{l_c,g_c}}{\#L_c} \tag{1}$$

where $Coverage_c$ is the coverage for a specific class c, $\sum C_{l_c,g_c}$ is the sum of the instances where the local (l_c) and global (g_c) relations have at least one element in common for class c, and $\#L_c$ is the number of local instances that belongs to the class c. Figures 2 and 3 present values of $Coverage_c$ for each of the two datasets.

This analysis helps clearly identify classes with very low coverage (such as "reading") where the identified local relations are not likely to be influencing the classification task, versus classes with high coverage (such as "riding_a_bike") where significant global relations are present in most of the local instances. Figure 4 presents an example of how our method works for the class "riding_a_bike" in the *Action40* dataset.

It is shown that, based on the top concepts extracted from the first phase, ten local relations based on an instance of "riding a bike" and three global relations for the same label were selected as candidates. In the validation step, images from *Visual Genome*, which contain the concepts presented in the relation candidates, are used to verify which relation was learned by the model. In this case, only the "wheel is part of a bicycle" was identified as having been learned. As the relation learned is presented both locally and globally, we then define that the relation covers this case.

We might be tempted to say that for classes with low coverage, the model is likely to not have learned relations that are crucial for the classification tasks, and this might be used as a starting point to investigate how to better train the model for those classes, for example by looking at class imbalances or data augmentation techniques as well as knowledge injection mechanisms.

However, we are aware that the values for coverage might also depend on other factors. For example it might depend on how well the linear classifier identifies concepts that separates classes well. It might also be due to the quality of the concept extraction approach based on *Network Dissection*, which in turn might depend on the quality of the dataset. For a low-resolution dataset such as *CIFAR-10*, for example, we observe that only classes "dog" and "bird" have high coverage. This observation would need to be confirmed by conducting an ablation study to identify the quality of concepts and how they affect the overall pipeline, and this is an area for further investigation.

Furthermore, a deeper investigation would be required to identify the impact of different concept extraction techniques (both global and local), before reaching a conclusion on how well the extracted

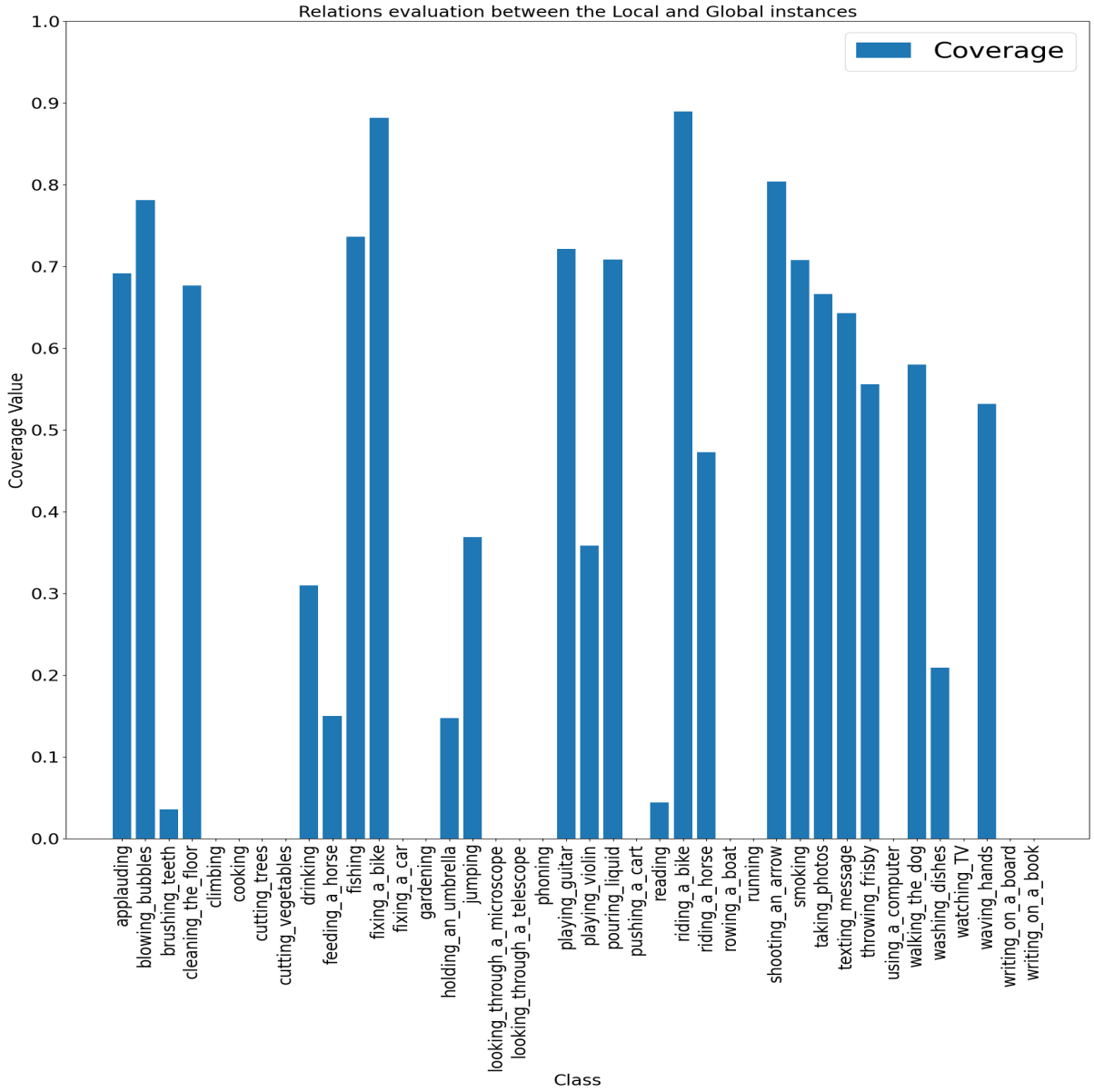

Figure 2: Coverage on *Action40*

relations are implicitly learned by the model.

5. Conclusion and Next Steps

Efforts to improve *CNNs*' transparency have insofar largely focused on visual cues to highlight what pixels in an image most influenced the prediction. While useful, these methods fall short in providing a truly understandable and human-friendly explanation. In response to this limitation, we developed an approach to extract and validate relations among disentangled concepts from *CNN* models trained on image classification tasks. Our approach combining concept extraction techniques and knowledge graphs paves the way towards a deeper understanding of trained *CNNs*' in terms of concepts and semantic relations among them, and therefore it has the potential to support human understanding of how these black-box models reach their decisions, and potentially improve the learning process.

We are aware that more investigation is required to reach this ambitious goal, and we have identified some limitations of our approach that we plan to extend in future work. One limitation of our

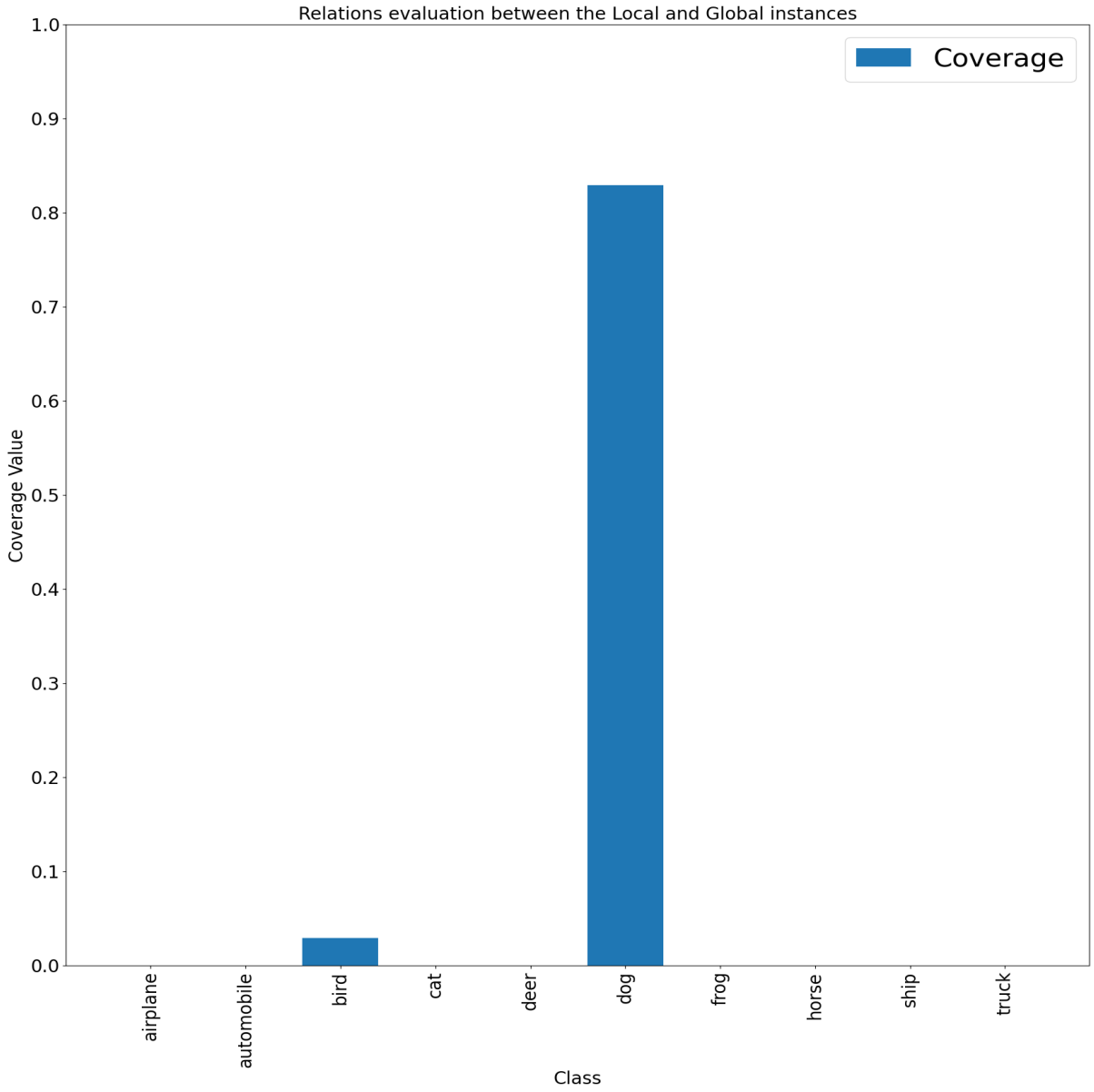

Figure 3: Coverage on *CIFAR-10*

method is its reliance on *Network Dissection*, which may not fully capture the complexities of human-understandable concepts. This could be addressed by relying on larger datasets with pixel-level concept labels, or by exploring alternative methods for disentangling concepts, such as Concept Activation Vectors [17] or *CLIP-Dissect* [30]. Additionally, there is the need to measure consistency between commonsense knowledge graph relations and the concepts derived from *Network Dissection* and *Visual Genome*, and define additional metrics that can help identify the suitability of different knowledge graphs. A broader comparison on the use of different image datasets and external knowledge graphs would be beneficial to analyse how the method proposed in this work can adapt to different domains and how the selection of validation data and knowledge graphs affects the results.

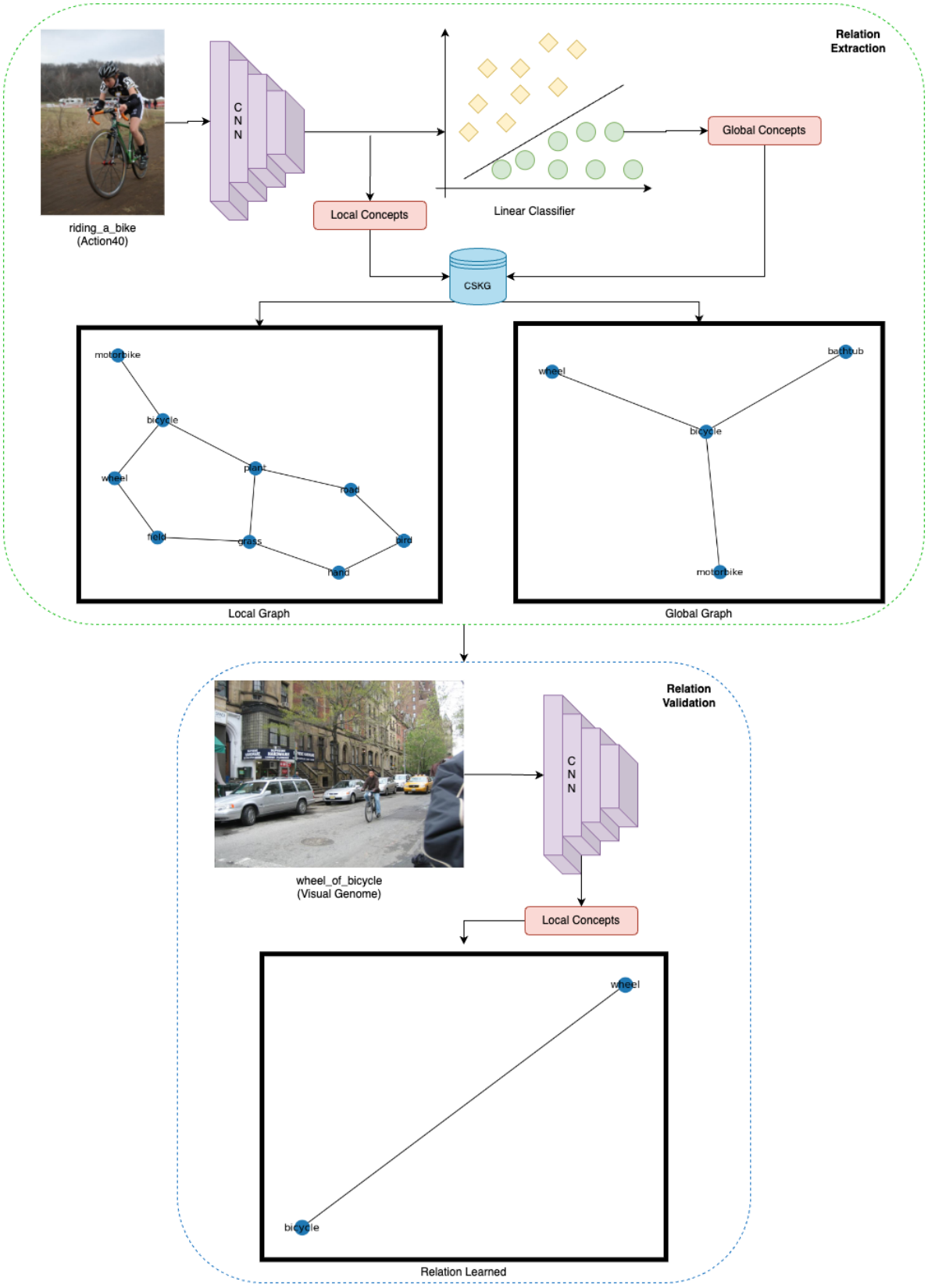

Figure 4: Relation Learned Example

Acknowledgments

Thanks to the financial support of Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223 and the Insight the SFI Research Centre for Data Analytics at Dublin City University under Grant No. SFI/12/RC/2289_P2.

References

[1] Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, MIT Press, Cambridge, MA, USA, 1998, p. 255–258.

[2] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, Cham, 2014, pp. 818–833. doi:10.1007/978-3-319-10590-1_53.

[3] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, 2017. doi:10.1109/ICCV.2017.74.

[4] G. Ras, N. Xie, M. van Gerven, D. Doran, Explainable deep learning: A field guide for the uninitiated 73 (2022). URL: https://doi.org/10.1613/jair.1.13200.

[5] E. Ferreira dos Santos, A. Mileo, From disentangled representation to concept ranking: Interpreting deep representations in image classification tasks, Springer, 2023. doi:10.1007/978-3-031-23618-1_22.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ArXiv abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[7] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. URL: https://api.semanticscholar.org/CorpusID:13756489.

[8] A. Mandal, S. Leavy, S. Little, Biased attention: Do vision transformers amplify gender bias more than convolutional neural networks?, 2023. URL: https://papers.bmvc2023.org/0629.pdf.

[9] D. Minh, H. X. Wang, Y. F. Li, T. N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artificial Intelligence Review (2021). doi:10.1007/s10462-021-10088-y.

[10] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, E. Baralis, Concept-based explainable artificial intelligence: A survey, 2023. URL: https://arxiv.org/abs/2312.12936.

[11] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on relation extraction: Recent advances and new frontiers (2024). doi:10.1145/3674501.

[12] H. Wang, K. Qin, R. Y. Zakari, G. Lu, J. Yin, Deep neural network-based relation extraction: an overview, Neural Computing and Applications (2022). doi:10.1007/s00521-021-06667-3.

[13] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, P. M. Atkinson, Explainable artificial intelligence: an analytical review (2021). doi:10.1002/widm.1424.

[14] G. Mutahar, T. Miller, Concept-based explanations using non-negative concept activation vectors and decision tree for cnn models, 2022. URL: https://arxiv.org/abs/2211.10807. arXiv:2211.10807.

[15] B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection 41 (2019) 2131–2145. doi:10.1109/TPAMI.2018.2858759.

[16] Q. Zhang, Y. Yang, H. Ma, Y. Wu, Interpreting cnns via decision trees, Los Alamitos, CA, USA, 2019, pp. 6254–6263. doi:10.1109/CVPR.2019.00642.

[17] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. URL: https://arxiv.org/abs/1711.11279. arXiv:1711.11279.

[18] X. Zou, A survey on application of knowledge graph, Journal of Physics: Conference Series (2020). doi:10.1088/1742-6596/1487/1/012016.

[19] S. Choudhary, T. Luthra, A. Mittal, R. Singh, A survey of knowledge graph embedding and their applications, 2021. URL: https://arxiv.org/abs/2107.07842. arXiv:2107.07842.

[20] L. Zhong, J. Wu, Q. Li, H. Peng, X. Wu, A comprehensive survey on automatic knowledge graph construction, ACM Comput. Surv. 56 (2023). URL: https://doi.org/10.1145/3618295.

[21] F. Ilievski, P. Szekely, B. Zhang, Cskg: The commonsense knowledge graph, 2021. URL: https://arxiv.org/abs/2012.11490. arXiv:2012.11490.

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, L. Fei-Fei, Visual genome: Connecting language and vision using crowdsourced dense image annotations 123 (2017). doi:10.1007/s11263-016-0981-7.

[23] R. T. Icarte, J. A. Baier, C. Ruz, A. Soto, How a general-purpose commonsense ontology can improve performance of learning-based image retrieval, 2017, pp. 1283–1289. doi:10.24963/ijcai.2017/178.

[24] I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey 302 (2022). doi:10.1016/j.artint.2021.103627.

[25] N. Maillot, M. Thonnat, Ontology based complex object recognition, Image Vis. Comput. 26 (2008) 102–113. URL: https://api.semanticscholar.org/CorpusID:11507373.

[26] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, 2011, pp. 1331–1338. doi:10.1109/ICCV.2011.6126386.

[27] A. Krizhevsky, Learning multiple layers of features from tiny images, 2009. URL: https://api.semanticscholar.org/CorpusID:18268744.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition (2015) 770–778. URL: https://api.semanticscholar.org/CorpusID:206594692.

[29] M. U. Hadi, Q. A. Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, S. Z. Hassan, M. Shoman, J. Wu, S. Mirjalili, M. Shah, Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects (2024). doi:10.36227/techrxiv.23589741.v7.

[30] T. P. Oikarinen, T.-W. Weng, Clip-dissect: Automatic description of neuron representations in deep vision networks (2022). URL: https://api.semanticscholar.org/CorpusID:248376976.