Causal Concept Identification in Open World Environments

Anonymous submission

Abstract

The ability to continually discover novel concepts is a core task in open world learning. For classical learning tasks new samples might be identified via manual labeling. Since this is a labor intensive task paper we propose to utilize causal information for doing so. Image data provides the ability to directly observe the physical real world appearance of concepts. However, the information presented in images is usually of a noisy and unstructured nature. In this position paper we propose to leverage causal information to structure and causally connect visual representations. Specifically, we discuss the possibilities of using causal models as a knowledge source for identifying novel concepts in the visual domain.

Overview. Section (1) motivates continuous concept discovery using causal mechanisms. Section (2) outlines a path to continually advance the discovery of visual concepts using causal structures. Section (3) outlines practical issues encountered in (1) and (2) and discusses future steps.

1 Motivation

Modern machine learning systems need to process large amounts of annotated image data to identify visual concepts. While the resulting models achieve impressive results and push the limits of the field, they lack human curiosity and fall short in their ability to perform lifelong learning. One drawback of such approaches is the necessity to provide a supervision signal for the image data. While modern approaches utilize large amounts of data available from the internet, the trained models fall short on training data for niche domains and might adapt harmful biases from the data. Continual learning approaches are interested in *continually discovering novel concepts* to help machine learning models improve and adapt to new environments. A key challenge in this regard is to provide sufficient amounts of accurately annotated data. For this purpose the field of causality can help to provide the required supervision by leveraging the structure of causal mechanisms.

In the following sections, we lay out a possible path of connecting continual, open world, learning with causality and discuss the challenges encountered along the way. We highlight the importance of causal knowledge to identify and disentangle objects in the visual domain.

How do humans discover novel concepts? Machine learning models are usually trained by presenting them with randomly sampled image-label pairs. These samples usually do not have any connections besides sharing the same set of labels. Machine learning models are expected to identify visual concepts from ground up. This is in contrast to the way humans discover the world. When presented with novel concepts we usually start out with some initial knowledge. We relate novel concepts to preexisting knowledge and therefore *continually* advance our understanding of the world (Chen and Liu 2018; Flesch et al. 2022).

Why do we want causal concept discovery? Causality is concerned with identifying processes that are underlying real world observations. As such, structural causal models (SCM) are designed to model the causal relationship between different concepts. Whenever we detect some concepts of a given SCM, we can utilize the graph structure to infer the presence of other, possibly unknown, concepts. Definitions of a 'concept' might vary, depending on the specific use case. In this paper we will use the term to capture the range from low-level features, such as color or texture, to complex composed entities, e.g. cities. Assume that our system is able to detect matches and matchboxes, but has not yet seen a flame. However, we might have read about the causal effect of sliding matches along the striking surface of the matchbox to light them. We can express this knowledge in a causal graph without having to provide image samples. Now, whenever we observe the interaction of striking matches on a matchbox, we can infer the presence of a flame and in consequence train our vision model on the new concept. This approach helps us in two ways: (1) Since the causal model specifies the exact conditions under which a particular concept is to be expected, we can actively steer our discovery process towards those instances and discover concepts more efficiently. (2) Secondly, causality helps us with disentangling concepts and ruling out confounding factors. While some objects might only appear in strongly correlated settings we know the *true causal factors* from the SCM.

2 Causal Concept Discovery

Discovering concepts under causal supervision provides a way to incrementally discover novel concepts. For every observation, we search for an SCM that contains the already

known concepts and follow the causal paths to arrive at novel concepts. Each of those discovery step allows us to align more causal graphs without observations and continually broadens the scope of our models.

Structural Causal Models. As we use SCMs to *ground* and discover novel concepts in the visual domain, we follow the Pearlian notion of Causality (Pearl 2009). An SCM is defined as a 4-tuple $\mathcal{M} := \langle \mathbf{U}, \mathbf{V}, \mathcal{F}, p(\mathbf{U}) \rangle$ where the so-called structural equations $v_i \leftarrow f_i(\mathrm{pa}_i, u_i) \in \mathcal{F}$ assign values to the respective endogenous variables $V_i \in \mathbf{V}$ based on the values of their parents $\mathrm{Pa}_i \subseteq \mathbf{V} \setminus V_i$ and the values of their respective exogenous variables $\mathbf{U}_i \subseteq \mathbf{U}$. In particular, any SCM induces a *causal graph* which represents the causal structure from causes to effects.

Bootstrapping. SCMs are specifically tailored to represent information about causal systems. However, in practice we may not be able to explicitly provide a list of concepts C learned by the model and might even encounter catastrophic forgetting of already learned concepts (French 1999). This poses a practical problem as it introduces additional uncertainty to our system. For our theoretical discussion, we assume for now that we can reliably detect all concepts that we have already discovered during our process.

For starting up our discovery process, we assume an initial set of known concepts C_0 to be given, which can be reliably detected. This intial knowledge might come from training on manually annotated data sets or other pre-trained models (Lin et al. 2014; Krizhevsky, Sutskever, and Hinton 2012; Minderer et al. 2022). Additionally we assume to be given a set of SCMs which encode our causal knowledge about the world.

Boundaries of Discovery. Open world settings provide an infinite stream of concepts to discover (Chen and Liu 2018). Like for human curiosity, we are not interested in learning random concepts, but in utilizing our existing causal knowledge to efficiently discover concepts that stay close to our already existing knowledge. As such, we define the *causal frontier* as the set of SCMs that contain at least one concept of C_i. Importantly this gives us a method to a priori determine which concepts can and can not be discovered. Given the set of initial concepts C_0 and the set of given causal graphs, it follows that we can only discover those concepts for which we can find a chain of causal graphs, such that for any two adjacent SCMs we get a non-empty overlap between their set of variables.

Discovering concepts. At this point in our process we can start to advance our causal frontier by continually learning new concepts from observations. As a first step, we identify all known concepts that are present in a new observed image and select the causal graphs that contain those concepts. However, discovering that a concept is contained in the set of endogenous variables of a causal graph does not suffice to infer the presence of the causal system. Some common concepts, such as color, might appear as parameters in many causal graphs. Apples, for example, are typically colored green or red. As such, the concept of color *parameterizes* the observed object. However, it is not suited to infer the type of an object, as being colored red or green does not make an object an apple. Detecting the typical 'apple-like shape' however would be a strong indicator for the concept. Therefore, we are interested in SCMs for which we discover indicator variables or *actual causes* (Halpern 2016) that are necessary to infer the presence of a causal system.

Another important insight is the fact that some causal structures may only be identifiable with the help of interventions (Pearl 2009; Bareinboim et al. 2022). Consider, for example, a scenario of two independent variables, A and B, whose appearance are determined by a third variable C. As a consequence, A and B are either present at the same time or not, and we have no way to disentangle them. Causality can help to detect such situations. In consequence we can actively intervene on the system and identify the individual concepts.

3 Challenges and Future Steps

In the previous sections we outlined the high-level idea of causally guided concept discovery. For this section we now continue to discuss the challenges that may arise in practice.

Challenge 1: Identifying causal paths. One problem with identifying unknown variables from SCMs is the fact that we may not know how causes and effects are interacting in the real world. In our initial example of lighting a match we might follow the physical process of the objects interacting. For other examples we might assume to only observe sparse changes. While we can continue to come up with more heuristics for specific problems, we have to recognize that the current formalizations of causality are not well suited to trace causal effects in their underlying systems.

Challenge 2: Abstracting concepts. Another challenge towards identifying unknown concepts from SCMs arises from the fact that causal systems are often modeled using high-level relations. Because of that we might encounter several low-level entities on the way from cause to effect the are not modeled in the SCM. In order to be able to identify these concepts, we need to consider abstractions and refinements of SCMs (Beckers and Halpern 2019; Rubenstein et al. 2017). This means, that we have to come up with ways of identifying intermediate concepts or refine the given SCMs to better reflect the abstraction level of our observations.

Summary and Outlook: In this position paper we highlighted the strengths and challenges of continuous causal concept discovery. We presented ways on how to leverage causal structures to guide concept discovery and identify novel concepts. While we primarily focused on the application of causal knowledge to discover open world concepts, the inverse problem of inferring causal knowledge from open world settings is also still to be discussed. Identifying relevant concepts and connecting them in a way that leads to meaningful causal concepts poses a challenging problem on its own that is yet to be solved. For future applications, we might consider combined approaches that discover visual concepts via causal guidance while simultaneously refining their causal knowledge using observations.

References

Bareinboim, E.; Correa, J. D.; Ibeling, D.; and Icard, T. 2022. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, 507–556.

Beckers, S.; and Halpern, J. Y. 2019. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, 2678–2685.

Chen, Z.; and Liu, B. 2018. Lifelong machine learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 12(3): 1–207.

Flesch, T.; Nagy, D. G.; Saxe, A.; and Summerfield, C. 2022. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Halpern, J. Y. 2016. *Actual causality*. MiT Press.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *European Conference on Computer Vision (ECCV)*. Springer International Publishing.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. *ECCV*.

Pearl, J. 2009. *Causality*. Cambridge university press.

Rubenstein, P. K.; Weichwald, S.; Bongers, S.; Mooij, J. M.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*.