Context-Dependent Deep Learning

Apprentissage en Profondeur Dépendant du Contexte

Roy M. Turner[1], Cynthia Loftin[2], Alexander Revello[1], Logan R. Kline[3], Meredith A. Lewis[4], and Salimeh Yasaei Sekeh[1]

[1] School of Computing and Information Science, University of Maine, Orono, Maine, USA, rturner@maine.edu
[2] U.S. Geological Survey, Maine Cooperative Fish and wildlife Research Unit, Orono, Maine, USA, cynthia.loftin@maine.edu
[3] Ecology and Environmental Science, University of Maine, Orono, Maine, USA, logan.kline@maine.edu
[4] Ecology and Environmental Science, University of Maine, Orono, Maine, USA, meredith.a.lewis@maine.edu

ABSTRACT. Explicitly representing an agent's context has been shown to have many benefits, which should also apply to machine learning. In this paper, we describe an approach to do this called context-dependent deep learning (CDDL), which is based on earlier work in context-mediated behavior (CMB) that uses contextual schemas (c-schemas) to represent classes of situations along with knowledge useful in them. These c-schemas are then recalled and guide reasoning in the corresponding contexts. CDDL stores knowledge about deep neural network structure and weights in c-schemas, which allows context-specific learning. Our work is being developed in the domain of seabird detection in aerial images of islands for use by biologists.

RÉSUMÉ. La représentation explicite du contexte d'un agent a montré avoir de nombreux avantages, qui devraient aussi pouvoir s'appliquer à l'apprentissage symbolique (CDDL). Dans ce papier nous décrivons une approche pour faire ceci qui est appelée apprentissage en profondeur dépendant du contexte. Cette approche est basée sur des travaux antérieurs sur le comportement médiatisé par le contexte qui utilise des schémas contextuels (c-schémas) pour représenter des classes de représentations avec les connaissances utiles pour elles. Ces c-schémas sont rappelés pour guider le raisonnement dans les contextes correspondant. CDDL stocke les connaissances sur la structure du réseau neuronal profond et des pondérations dans les c-schémas, connaissances qui permettent un apprentissage contextualisé spécifique. Ce travail est réalisé dans le domaine de la détection des oiseaux de mer sur des images aériennes d'iles pour des biologistes.

KEYWORDS. Deep learning, neural networks, context-mediated behavior, object detection, image recognition.

MOTS-CLÉS. Apprentissage,profond, réseaux de neurones, comportement médiatisé par le contexte, détection d'objets, reconnaissance d'image.

Over the past two decades, researchers in the context community have shown the importance of explicitly representing (modeling) and using context and contextual knowledge in a wide range of areas, including natural language processing, problem solving, and handling unanticipated events.[1] Simultaneously, machine learning has progressed rapidly in many of the same areas, often outpacing traditional symbolic approaches. To date there has been little cross-pollination of machine-learning and symbolic approaches, yet context is as important for machine learning as it is for any other task.

A machine learning system working in many different contexts necessarily must learn contextual and task features and relationships simultaneously. For example, a deep learning neural network trained to identify objects in images has to learn what objects look like in many different contexts: alone, in a cluttered environment, under different lighting conditions, etc. The system develops only an implicit model of context that is spread across the network's weights. However, explicitly representing context has numerous benefits, including helping understand how context affects tasks, allowing reasoning about contexts as first-class objects to facilitate interpretation and acquiring

[1] See, e.g., the CONTEXT (International and Interdisciplinary Conference on Modeling and Using Context) conference series as well this journal.

knowledge from humans, and avoiding redundant work by reasoning *within* rather than repeatedly *about* the context.

A better approach is to separate the problem of identifying the task context from the problem of learning or performing tasks within it. The first problem can be addressed by existing context-sensitive reasoning approaches, such as our own context-mediated behavior (CMB) [11]. Then a machine learning system can then be tailored to the particular needs of tasks in that context. Since the work of identifying the context is done by another program, the machine learning system should learn faster, from fewer examples, and require fewer resources (e.g., number of weights, amount of processing time, etc.).

Indeed, a pilot study some years ago showed just this, at least for very simple neural networks [1]. Neural networks were used to assess the depth of an autonomous underwater vehicle (AUV) both alone and in conjunction with a context manager. The context manager was responsible for diagnosing the situation as an instance of a known context, then used information about the context to instantiate a network with appropriate structure and weights learned in past occurrences of the context. Experiments showed that context-specific networks can be smaller yet still produce fewer errors than context-independent networks and that training time for these networks can grow more slowly as the number of different contexts increases. Work by Stein & Gonzalez (e.g., [10]) has shown that applying context-based reasoning to a neuroevolutionary, learning-from-observation task can significantly improve performance by allowing the system to learn how to behave in contexts identified by humans as pertaining to parts of the task.

In this paper, we discuss an ongoing project to extend our early work in deep learning to more difficult tasks requiring much more complex networks. The goal of the project is to develop a general reasoning mechanism, which we call *context-dependent deep learning* (CDDL), to be used with deep learning neural networks in a variety of domains. Our initial domain is recognizing different types of nesting birds on islands off the coast of Maine (USA) from aerial imagery.

Detecting birds in aerial images

Our work is part of a project that involves acquiring and analyzing aerial imagery of 274 of Maine's offshore islands to identify and count nesting seabirds. This is important because seabird populations are sensitive to the dynamics of their forage-fish prey, and so are indicators of the health of the marine and coastal environments [2]. Biologists survey these populations during the breeding seasons, but surveys are challenging due to the remoteness of the islands. However, aerial (plane-based) images are available both historically and from current surveys. Figure 1 shows an image of a small portion of one island with gulls located within a context of cobble.

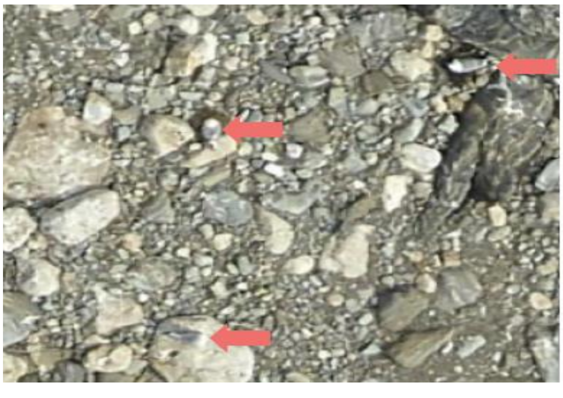

Figure 1. *A portion of a representative image context, with Herring Gulls (Larus argentatus) indicated.*

Processing images manually is labor-intensive, time-consuming, and error-prone. For example, processing imagery for one of the islands involved identifying 2500+ birds and required approximately 14 hours each for two people. Given the large number of islands of interest and the desire to survey the islands on a regular basis, human image analysis is impractical. Fortunately, there are many different approaches to computer-based image recognition, the most promising of which is deep learning with *convolutional neural networks* (CNNs) [4]. CNNs are patterned loosely after mammalian visual systems, in which a series of functions (convolutions) is performed on the input image to detect increasingly abstract features until finally entire objects are detected. Each neuron in one of the convolutional layers has a small receptive field in the preceding layer it responds to in particular ways. Many convolutional layers are stacked together, with the resulting highly-abstract features usually passed on to fully-connected layers at the end to recognize objects.

Our domain requires recognizing multiple objects in an image and providing locations and bounding boxes for them (*image segmentation* and *localization*, resp.). Deep CNNs that can do this include the Faster R-CNN (Region CNN) [9], YOLO (You Only Look Once) [8], and SSD (Single Shot Multibox Detector) [5]. Both YOLO and SSD are faster (during recognition) than Faster R-CNN, and YOLO has been shown to be effective for recognizing small objects (relative to the image size) in aerial photographs [6]. At least one study [7] found that YOLO was slightly more accurate than SSD, although susceptible to false positives when the objects were of varying sizes. In our domain, size variability is small, and so we are using YOLO. Figure 2 shows an example of YOLO's output for our domain.

Unfortunately, the depth of these networks and total number of parameters to be trained (e.g., YOLO v3 has 53 convolutional layers and roughly 65 million weights) result in long training times and require a large number of labeled images as training examples. For example, one of our recent training sessions with YOLO resulted in only a 60% recognition success after a week of training on a state-of-the-art, shared multi-GPU cluster capable of several petaFLOPs (quadrillions of floating point operations per second).

Learning in context

Task complexity can increase training time and network size (depth and number of weights). Our earlier pilot study [1] showed that, at least for simple neural networks, removing the need for context recognition by the network decreases the network size required in any given context while preserving performance quality. The study also suggested that allowing networks with different structure, weights, or both, to be trained separately in different contexts can actually increase the networks' accuracy.

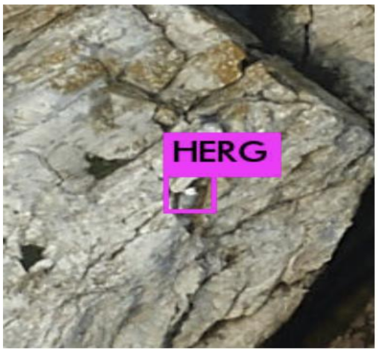

Figure 2. *Example output of YOLO for our domain, with one Herring Gull labeled.*

Using a similar approach for more complex networks in our domain promises several benefits. First, learning in context will reduce the network complexity and training time needed in any particular context; a network for detecting gulls among rocks and one for detecting them in grass, for example, are each likely to be simpler and easier to train than a single network that has to learn to recognize them in both contexts.

Second, learning in context will allow effective allocation of training time and processing. For example, objects can be more difficult to detect in some contexts than others. A nesting gull is *much* easier to detect in grass than among rocks, and so it makes sense to use a deeper network or more training in the rocky context. With a single network that does not explicitly recognize the context, this is not possible. Similarly, if objects are known not to occur in some contexts, then there is no need to look for them there. For example, there is no need to look for nesting herons in rocky or grassy areas, nor to look for nesting gulls in trees. This will free up training time to look for the birds that *are* expected.

Third, by recognizing the context, training examples can be used more effectively. For example, one species of cormorant (the Great Cormorant, *Phalacrocorax carbo*) has been found on only a few islands in Maine, whereas the Double-crested Cormorant (*P. auritus*) is much more widespread. A network trained on images of cormorants drawn from across all islands may not learn to recognize the distinction between the two, whereas if the known occurrence of *P. carbo* is taken into account to separate the examples, the networks trained for those islands are more likely to be able to detect both species.

In our domain, as in many others, contextual knowledge can be readily acquired from human experts. Biologists know a great deal about where in general one is likely to find or not find nests of different bird species, as well as the relative difficulty of identifying birds in different kinds of terrain. While this knowledge cannot easily be given to a neural network—human knowledge is symbolic, while a neural network’s is sub-symbolic—it can easily be used by a symbolic AI program tasked with identifying the context.

Context-dependent deep learning

The overall CDDL process is shown in Figure 3. As in context-mediated behavior, a context manager (ConMan) observes the world to find *contextual schemas* (c-schemas) that most closely match the current situation. A c-schema is a frame-like knowledge structure that contains facts about the context represented (descriptive knowledge) as well as knowledge about how tasks should be performed in the context (prescriptive knowledge). A more complete definition and description of c-schemas can be found elsewhere (e.g., [11]). In our current domain, the world consists of the image being processed and any information added as annotations by humans (e.g., the type of camera used, the weather, where the image was taken, etc.). Finding the most appropriate c-schemas is a diagnostic task that begins with candidate c-schemas being "evoked" by features of the world, then critically compared with each other and the situation.

A situation may be an instance of multiple contexts (e.g., "rocky island", "*P. carbo* expected", "cloudy", etc.). In such cases, ConMan's diagnosis will consist of multiple relevant c-schemas that are merged into a coherent representation of the combined context called the "Lens" (since it is how the system views the world). The Lens is essentially a c-schema created on the fly from other relevant c-schemas to represent the current context. In past work in intelligent agent control applications, knowledge contained in the Lens was used to describe the situation and to prescribe how to behave. In this work, it will provide information needed to instantiate a neural network to recognize/learn to recognize objects in the context of the image.

Some features that we have identified as useful for characterizing contexts in this domain include the terrain (rocky, grassy, etc.), location (latitude/longitude, specific island, etc.), altitude of the camera, weather (sunny, cloudy), time of day/year, type of camera, and whether the island is inhabited. In a planned longer work to appear soon, we will characterize the feature space more completely.

In some contexts, entire prior neural networks, including weights, should be provided; in others, the Lens may suggest creating a new neural network base, with its topology and kind of layers based on the system's past experience and knowledge from experts. Hyperparameters appropriate for the context can also be provided, for example learning rate and other properties of training, such as the number of epochs expected to be needed (related to the task's expected difficulty in the context).

When a task is complete or the context changes, ConMan updates the Lens' component c-schemas with anything learned (e.g., changed weights). This updating allows context-specific learning, which can then be used in the future when similar contexts are encountered.

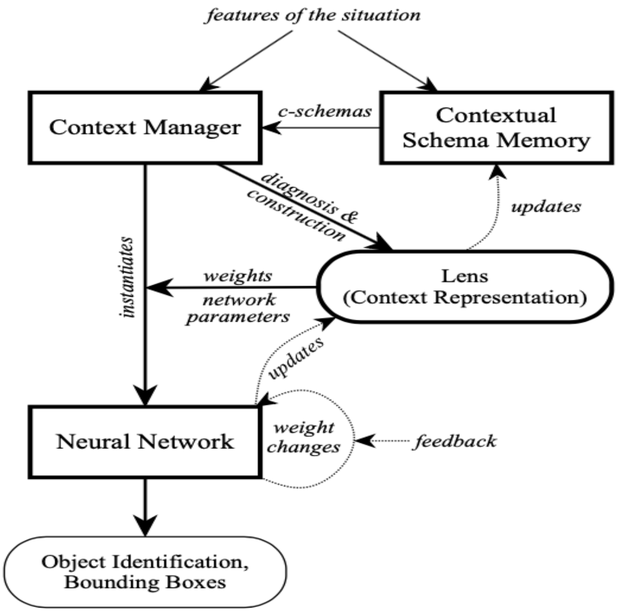

Figure 3. *Overview of the CDDL process.*

While we expect some contexts to be identified by human experts initially, the goal is to have ConMan learn contexts (and thus, contextual schemas) based on its own experience. The kind of memory used in our past work directly supports simple inductive learning based on comparing and differentiating between existing contextual schemas and new instances of problem solving. We expect to augment this approach with other learning techniques as the work progresses (e.g., deep learning to detect contextual features important for successful identification of birds that may indicate new contexts the system should learn and use).

Planned experiments

Experiments are planned for the summer of 2021 to determine if CDDL is useful in our domain. The control for these experiments will be the current version of YOLO now being trained without regard to image context. This version will be compared to smaller networks trained on image subsets drawn from similar contexts (e.g., grassy area versus rocky shore, weather condition, etc.). Training time for the control network will be compared to total training time for the smaller networks, and performance on test images will be compared.

Data will also be gathered about the overhead of context-switching. A complete CDDL system will almost certainly be presented images without regard to their context, especially as new images are acquired over time. The CDDL system will assess the context of each image and, when the context changes, store current network properties and instantiate a new network (or at least load different weights). We will measure the time and space requirements when the context changes due to images being presented in a random order, which will also provide data about memory/disk requirements for CDDL in this domain. (The time needed to notice a context change and find new c-schemas, etc., will be assessed in future once ConMan is implemented). We are exploring compression techniques one of us has developed (e.g., [3]) to reduce the likely substantial space requirements(e.g., weights for one pre-trained version of YOLO v3 require ~250 MB.). Other experiments will compare the performance of the networks under different frequencies of errors in context diagnosis.

Conclusion and future work

CDDL involves off-loading context recognition from neural networks when the domain requires tasks to be done across multiple contexts, which should reduce size and learning time as well as promote in-context learning. The approach will also allow existing contextual knowledge to be acquired from humans, rather the network having to learn it from examples.

Work on CDDL is in an early stage, with our effort so far focusing on designing the overall process. We have implemented and are training a version of YOLO on our high-performance GPU cluster and have begun delineating some of the situational features and network properties important to represent in CDDL. In the near future, experiments will be performed to evaluate our approach.

Assuming positive experimental results, a full version of CDDL will be implemented. It will be a testbed for future research in this and other domains as well as being made available on our cluster for use by the biologists, using a web-based interface recently completed as a student capstone project. Further in the future, we will investigate using deep learning for aspects of context diagnosis itself.

Acknowledgments

We thank the University of Maine Research Reinvestment Fund and AI Initiative for funding this study. At the time of publication, data sets used in this study were not available or have limited availability (i.e., ongoing work, proprietary, or sensitive). We thank A. Gonzalez and Y. Shang for their reviews of the manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The high-performance cluster used in this work was funded by a grant from the National Science Foundation (Major Research Instrumentation, #1919478). We also thank the Advanced Computing Group, and especially Steven Cousins, for cluster support and assistance. This study was performed under the auspices of UMaine protocol # A2020-03-01.

References

[1] Arritt, R. P., and Turner, R. M. Context-specific weights for a neural network. In *Proceedings of the Fourth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'03)* (Stanford, CA, 2003), Springer, New York, pp. 29–39.

[2] Furness, R. W., and Camphuysen, K. Seabirds as monitors of the marine environment. *iceS Journal of Marine Science* 54, 4 (1997), 726–737.

[3] Ganesh, M. R., Blanchard, D., Corso, J. J., and Sekeh, S. Y. Slimming neural networks using adaptive connectivity scores. arXiv preprint arXiv:2006.12463 (2020).

[4] LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Backpropagation applied to handwritten Zip Code recognition. *Neural Computation* 1, 4 (1989), 541–551.

[5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. SSD: Single Shot MultiBox Detector. arXiv:1512.02325 [cs] 9905 (2016), 21–37.

[6] Lu, J., Ma, C., Li, L., Xing, X., Zhang, Y., Wang, Z., and Xu, J. A vehicle detection method for aerial image based on YOLO. *Journal of Computer and Communications* 6, 11 (Nov. 2018), 98–107.

[7] Morera, Á., Sánchez, Á., Moreno, A. B., Sappa, Á. D., and Vélez, J. F. SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities. *Sensors* (Basel, Switzerland) 20, 16 (Aug. 2020).

[8] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You Only Look Once: Unified, real-time object detection. In 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788.

[9] Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2016).

[10] Stein, G., and Gonzalez, A. Learning in context: enhancing machine learning with context-based reasoning. *Applied Intelligence* 41:709–724 (2014).

[11] Turner, R. M. Context-mediated behavior. Modeling and Using Context 17, 1 (March 2017).