November, 2020

Dan Rosenbaum - Research Statement

In my research I aim to study computational models of perception and 3D scene understanding, and in particular generative approaches that model perception as an inverse problem. Popular deep learning methods that implement perception as feed-forward one-shot prediction, trained end-to-end, have recently led to very successful applications in simple problems like object recognition and detection, however these methods are facing major challenges when dealing with more general 3D scene understanding problems that require more flexible and data efficient solutions. To tackle these challenges, I am interested in developing **generative perception** methods that combine the excellent optimisation power of deep learning with more flexible, generalisable and data efficient models for perception of 3D scenes. I believe that improving these methods can translate to substantial and positive impact on the world, by leading to a better understanding of human intelligence, to technological advances in robotics, and to better computational imaging methods that can ultimately lead to scientific breakthroughs.

Background

In recent years there has been huge progress in methods that solve tasks like visual object recognition and detection, visual 3D scene mapping and navigation, and solving almost any pre-defined goal using reinforcement learning (RL) with an appropriate reward function. Most of this progress if fueled by methods following the so-called ‘deep learning’ approach which roughly consists of the following steps:

- A task / goal is defined along with a corresponding metric function.
- Supervised data is collected, containing input and output pairs along with their corresponding values of the metric function.
- A deep neural network is trained to fit the input - output pairs and maximise the metric.

This is true not only for supervised learning methods like object recognition where large datasets of images and their corresponding true object category are used for training, but also for reinforcement learning methods where the data is collected in an online fashion using a policy derived from the same neural network that is being trained or a separate exploration policy.

While in both cases some care is required in turning the actual metric to a suitable function like a differentiable loss for supervised learning or a less sparse reward function in reinforcement learning, deep learning is an end-to-end learning approach. It rests on the assumption that in order to obtain generalisation on new test data, all that is needed is to fit the input -> output function and maximise the metric for a large amount of training data. This is essentially a discriminative approach taken to the extreme.

In my research I challenge this fully discriminative approach. While its main advantage is that the final metric is being optimised directly, it also comes with serious disadvantages including

the need for very large amounts of data, and lack of flexibility for changes at test time. Therefore, I seek to develop methods that combine the excellent optimisation power of deep learning with a more generative approach, which is more data efficient, more generalisable, and allows more control of the model, making it more adaptive to changes at test time and to different prior knowledge.

In contrast to the deep learning approach which focuses on end-to-end training of a predefined task, in the generative approach models are trained to capture as much information and statistical structure in the data as possible. While posing a much more challenging problem for training, these models are encouraged to not throw away any information about the data and can therefore be used to solve various tasks at test time through probabilistic inference methods.

During my PhD, I studied generative methods for low level vision tasks like image restoration, optical flow and depth estimation. In Rosenbaum Zoran and Weiss (2013) we show how a statistical model of optical flow in image patches can be used as a prior and lead to better optical flow estimation between two images. In Rosenbaum and Weiss (2016a, 2016b) we show similar results using models of image warping noise and depth estimation. In Rosenbaum and Weiss (2015) we show that learning an inference network on top of a pre-trained statistical model of natural images, essentially combining a generative approach with deep learning methods, leads to state-of-the-art and fast image restoration. This method demonstrates the ability to get the best of both worlds by achieving the inference speed of neural networks and the adaptivity of the generative approach.

Generative perception

Perception can be defined as a mapping from sensory input (observations) into a representation that is useful for some downstream task. Thinking of visual perception for example, observed images need to be processed into a representation that makes them useful for navigation or any other task in the 3D scene. In recent years I've been focusing my research on developing methods that separate the perception problem and train perception models independently of the desired task. The main hypothesis is that this approach is more data efficient as it is guided by predicting the full high-dimensional observation, in contrast to methods that are trained using a specific metric for a task, and essentially throw away many bits of information.

In my research I have explored different ways a learned model can represent a 3D scene from visual observations. Together with Ali Eslami, Danilo Rezende and other colleagues at DeepMind, we have developed the Generative Query Network (GQN) model (Eslami et al. 2018). This model is trained to predict images of different viewing angles from a 3D scene, conditioned on a set of reference views. The model therefore is forced to represent the observed reference images in a way that captures the full structure of the 3D scene. In the paper, we show how this implicit representation can accelerate the learning of downstream tasks like controlling a robot arm.

Understanding the difference between 3D scenes representations obtained by different training approaches, is a central goal of my research. In Rosenbaum et al. (2018) we study the question of camera localisation in new 3D scenes, that were not previously seen by the model in training time. We observe that inferring the camera position using a model that was trained for perception, results in a better calibrated uncertainty landscape than using a model that was directly trained for the camera localisation task. This shows that in some cases, even if large amounts of supervised data is available, end-to-end training focusing on a specific task can result in inferior results compared to a model that is trained on the full perception problem, and therefore is forced to capture more structure in the data.

Dealing with uncertainty is a core aspect of perception, since observations in most cases contain partial information and are noisy. A big part of my research therefore focuses on developing adaptive models that can be conditioned on different amounts of information, and can deal with different levels of uncertainty. In Garnelo et al. (2018a, 2018b) and in followup work (Le et al. 2018 and Kim et al. 2019) we develop *Neural Processes*, training neural networks in adaptive conditioning settings like those in which Gaussian processes are used. The model provides a general framework for dealing with uncertainty, allowing a smooth transition between a prior model that is not conditioned on any data, and posterior models which can be conditioned on more and more data.

Future Perspectives

Recent years have seen the deep learning approach being successfully applied to more and more problems of various nature. During my time as a research scientist at DeepMind, working in what is probably the biggest deep learning lab in the world, I have been part of, and closely witnessed the many success stories of this approach, but I've also seen the limitations that it holds. Lately there has been a growing interest in understanding these limitations and finding ways to overcome them, combining ideas from different fields like classical computer vision, statistical signal processing and probabilistic graphical models. I believe that using deep learning more as a tool than as an approach is key to make machine learning more efficient and flexible. In particular I intend to contribute towards this goal through research in the following concrete topics.

Generative reasoning. As we show in Rosenbaum et al. (2018), training a model with a loss defined in observation space rather than label space can lead to better capturing of the underlying uncertainty landscape, which in turn can lead to more efficient and flexible models. Following up on this work I have started a collaboration with Peter Battaglia of DeepMind studying the characteristics of 3D scene representations when using loss functions that ground the uncertainty in observation space through a forward model. This is also closely related to the study of causality, where it is known that modeling data in the causal direction leads to modularity and better data efficiency. After organising a workshop on this topic in NeurIPS 2019 titled 'Perception as generative reasoning', we are now looking at fundamental aspects of this approach including the efficiency of learning symmetries. I plan to follow this direction by gathering empirical evidence comparing different training approaches. I believe that a

systematic study of learning via *generative reasoning* can lead to a better understanding of known phenomena in machine and human perception, and perhaps discover new ones.

Discrete representations. One of the holy grails of artificial intelligence (AI) research is to develop learning methods from which symbolic representations or *concepts* can emerge. Roughly, this can be defined as semantically meaningful discretisations in different parts of the representation. I believe that modeling perception as an inverse problem, and explicitly modeling the forward/causal direction can prove key to a meaningful discretisation of the representation, capturing semantics like objects, object types, events and concepts in general. In turn this can allow the usage of other discrete structures like graphs and attention mechanisms. In addition to how this discretisation can be achieved, there's the question of how it can be used to enhance artificial intelligence methods. I intend to pursue research on these two parallel questions, and towards this end, together with Volodymir Mnih of DeepMind, I have started to study them in the context of reinforcement learning. I envision this as a promising direction with the potential of making substantial progress in AI methods.

Embodied generative models. One important difference in the way machine learning models are trained compared to human learning is the distribution of training data. Whereas humans get a very structured and dependent stream of data, most machine learning methods rely on a set of i.i.d. training data. A child that is learning to perceive an object will not get a set of i.i.d. images of that object, but rather will try to move the object in different ways and try to understand the resulting changes in observation. I believe that an important aspect for achieving more efficient learning is having a model that can control the key variables in the data it is trained to capture. In Mellor et al. (2019) we show how using a model that can control a brush to recreate images can lead to interesting representations of these images in surprising ways. In a collaboration with Jay Mcclelland of Stanford University, we are studying questions around efficient learning of concepts through models that can control variables in the data. We hypothesise that this is an important factor for efficiently learning flexible and adaptive representations. I plan to follow this direction for 3D modeling where models can control objects in the scene in various ways.

Scientific imaging. There are several scientific fields that rely on computational processing of observations like data from telescopes for astronomy and data from microscopes for biology. These are usually extremely challenging perception problems that require careful dealing with uncertainty. Modeling perception as an inverse problem makes it easy to inject prior knowledge like the physical model of the microscope or a prior over the structure of molecules. I have started a collaboration with Olaf Ronneberger of the University of Freiburg and DeepMind applying the generative approach to the prediction of 3D structure of proteins. I intend to pursue this direction and I believe that the joint study of scientific imaging applications together with perception methods is doubly beneficial, as the application poses challenging settings to study perception, and progress in perception methods can lead to important scientific breakthroughs.

References

A. Eslami, D. J. Rezende, F. Besse, F. Viola, A. S. Morcos, M. Garnelo, A. Ruderman, A. A. Rusu, I. Danihelka, K. Gregor, D. P. Reichert, L. Buesing, T. Weber, O. Vinyals, D. Rosenbaum, N. Rabinowitz, H. King, C. Hillier, M. Botvinick, D. Wierstra, K. Kavukcuoglu, D. Hassabis. **Neural scene representation and rendering**, *Science*, 360(6394):1204–1210, 2018.

M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. Teh, D. J. Rezende, A. Eslami. **Conditional neural processes**, *International Conference on Machine Learning (ICML)*, 2018a.

M. Garnelo, J. Schwarz, D. Rosenbaum, F. Viola, D. J. Rezende, A. Eslami, Y. W. Teh. **Neural processes**, *Theoretical Foundations and Applications of Deep Generative Models Workshop, International Conference on Machine Learning (ICML)*, 2018b.

H. Kim, A. Mnih, J. Schwarz, M. Garnelo, A. Eslami, D. Rosenbaum, O. Vinyals, Y. W. Teh. **Attentive neural processes**, *International Conference on Learning Representations (ICLR),* 2019.

T. A. Le, H. Kim, M. Garnelo, D. Rosenbaum, J. Schwarz, Y. W. Teh. **Empirical evaluation of neural process objectives**, *Bayesian Deep Learning Workshop. Advances in Neural Information Processing Systems (NeuIPS)*, 2018.

J. F. Mellor, E. Park, Y. Ganin, I. Babuschkin, T. Kulkarni, D. Rosenbaum, A. Ballard, T. Weber, O. Vinyals, A. Eslami. **Unsupervised doodling and painting with improved spiral**, *Machine Learning for Creativity and Design Workshop. Advances in Neural Information Processing Systems (NeuIPS)*, 2019.

D. Rosenbaum, D. Zoran, Y. Weiss. **Learning the local statistics of optical flow**, *Advances in Neural Information Processing Systems (NeuIPS)*, 2013.

D. Rosenbaum, Y. Weiss. **The return of the gating network: combining generative models and discriminative training in natural image priors**, *Advances in Neural Information Processing Systems (NeuIPS)*, 2015.

D. Rosenbaum, Y. Weiss. **Beyond Brightness Constancy: Learning Noise Models for Optical Flow**, *arXiv preprint,* arXiv:1604.02815, 2016a.

D. Rosenbaum, Y. Weiss. **Statistics of RGBD Images**, *arXiv preprint,* arXiv:1604.02902, 2016b.

D. Rosenbaum, F. Besse, F. Viola, D. J. Rezende, A. Eslami. **Learning models for visual 3d localization with implicit mapping**, *Bayesian Deep Learning Workshop. Advances in Neural Information Processing Systems (NeuIPS)*, 2018.