Deep Learning for Visual and Virtual Worlds

Eleonora Vig, German Aerospace Center (DLR), Munich, Germany.

Recent years have seen a dramatic progress in the development of computer vision algorithms that rival and even surpass human abilities. These advances have huge implications for many of today's societal challenges, with much more to come. Behind the recent breakthroughs is deep learning, a machine learning method that aims at learning complex hierarchical visual representations "end-to-end", directly from data to optimally perform a certain visual task. Deep artificial neural networks have been consistently outperforming traditional, so-called "shallow" approaches on most classical vision problems, such as object recognition and detection, human action recognition, or semantic scene labelling. Novel applications (such as image captioning and image style transfer) as well as new record-breaking results on challenging benchmarks are reported regularly.

First, I will give a brief introduction to convolutional neural networks (CNNs), a fundamental data structure in deep learning for computer vision that has been inspired by the organisational principles of biological visual systems.
It has long been known that deep networks can learn arbitrarily complex feature hierarchies and decision functions, and recent advances in computing power have enabled us to build larger and larger networks with up to hundreds of millions free parameters.
However, these free parameters need to be tuned by a "training" process that presents examples and their desired outputs to the network, and a major challenge in building intelligent systems has now become the acquisition of sufficiently large, problem-specific datasets. Even worse, training often requires the time-consuming and expensive manual labelling of examples with the "ground truth".

In the second part of my presentation, I will discuss two aspects of this challenge and first steps towards their solution. A common limitation in applications of machine learning in general and deep learning in particular is the "dataset bias":
Typically, high-quality large-scale labeled training datasets are only available for very specific problems because it is either too costly or impractical to obtain such data for all possible real-world applications. The field of "domain adaptation" now may provide a solution to this challenge by training a model on a source data distribution so that it performs well on a different (insufficiently labelled) target data distribution. As an example, I will demonstrate how reusing and adapting object detectors pre-trained on general-purpose image datasets improved object detection and tracking performance on real-world videos of driving situations, which comprise a critical testbed for autonomous driving algorithms.

Another approach to overcome the annotated data scarcity is the use of algorithmically generated training data. I will describe a novel semi-automatic real-world cloning technique that can efficiently generate photo-realistic virtual worlds for all major semantic video analysis tasks, including semantic segmentation, depth estimation, and multi-target detection and tracking.

Finally, I will present an overview of ongoing work at the German Aerospace Center where we are developing deep learning algorithms for unconstrained aerial and satellite imagery. The problems range from aerial crowd monitoring, such as person counting, detection, and tracking at mass events, to multi-class object detection and road infrastructure mapping from airborne images for autonomous driving.