Using Deep Convolutional Neural Networks to guide the blind or visually impaired

Denis Shagalin[1], Lahav Barak[1], and Boris Glyazer[1]

[1] Department of Management, Bar Ilan University, Ramat Gan, 52900, Israel

Abstract

This chapter considers what it means to learn and navigate the world with limited or no vision. It investigates the limitations of blindness research underlying the spatial cognition of blind people. It discusses how technology can be in the form of an application installed on one's cellphone, using neural networks to help the visually impaired, or autopilots in convolutional neural networks (ResNet), and glasses with built-in cameras and inertial sensors that offer some needed guidance on the development of new spatial learning strategies and technological solutions that will ultimately have a significant positive impact on the independence and quality of life of this demographic.

1 Introduction

Not being able to see the world surrounding us is a terrifying thought; how often do we think about people with limited or no vision? How do people with such problems manage? Being blind or seeing poorly is not an easy task. The number of visually impaired and blind people is small - [2] about 300 million people worldwide. But according to alarming WHO data, that number is on a steady upward trend. There are estimates of an increase in the number of blind and visually impaired to 500 million

by 2030. It turns out that the search for and solution to this problem is largely a social plan. The brain's processing of visual information is a very important aspect of our daily lives, [1] so visually impaired people are at a disadvantage because the necessary information about the environment is not available to them. People who are limited in their ability to recognize the objects surrounding them depend on the people around them and auxiliary tools. For example, the help of a person accompanied, a guide dog, or a cane. Each auxiliary tool has certain drawbacks since a person must constantly be near the blind, the dog can become confused and takes a long time to learn, and the cane does not allow you to see approaching objects and limits your horizons—perception of surrounding objects. Therefore, we came up with the idea of studying, searching for, and offering an alternative set of modern technologies, as well as the use of convolutional neural networks (CNN) to transfer data collected from the environment by a machine that expresses to the Visually Impaired what it sees in front of it. This technology can be in the form of an application installed on one's cellphone.

2 Using Neural Networks to Help the Visually Impaired

In this article, we propose the use of [3] Convolutional Neural Networks (CNNs), Transfer Learning, and Partially Supervised Learning (SSL) to build a framework designed to help the visually impaired. It has low computational costs and, therefore, can be implemented on modern smartphones in conjunction with special glasses with a camera, autopilots, and sensors to collect additional information. The glasses camera can be used to automatically take pictures of the path ahead. They will then be immediately categorized, providing the user with near-instantaneous feedback. We also provide a dataset for training classifiers, including indoor and outdoor situations with different types of lighting, floors, and objects. Many different CNN architectures are evaluated as feature extractors and classifiers by fine-tuning weights pre-trained on a much larger dataset. A graph-based SSL method known as particle competition and collaboration is also used for classification, allowing user feedback to be considered without

retraining the underlying network. Classification accuracy of 92% and 80% is achieved in the proposed data set in the best-controlled scenario and the SSL scenario, respectively.

3 Using Autopilots in Convolutional Neural Networks (ResNet)

[2] The neural network does not understand what is shown in a photo or video from a camera. For her, it's just a set of shades from white to black, which are stored as color gradients. And to identify the video object, you need to process millions of pixels, distinguish them by color gradient, and remember correctly identified combinations. Neural networks need memory to store inputs, weights, and activation functions. Therefore, the requirements for computing resources for these systems are very high.

For the implementation of computer vision of autopilots of robotic devices, convolutional neural networks (ResNet) are most suitable. It is these networks that demonstrate the best indicators of accuracy and speed. However, their needs for computing resources are very high. For example, a 50-layer ResNet has about 26 million weights and computes 16 million forward activations. Using a 32-bit floating-point number to store each weight and activation would require about 168 MB. Additional memory is also needed to store input data, temporary values, and program instructions. Measuring memory usage when training ResNet-50 on a high-performance GPU showed that it requires more than 7.5 GB of local DRAM.

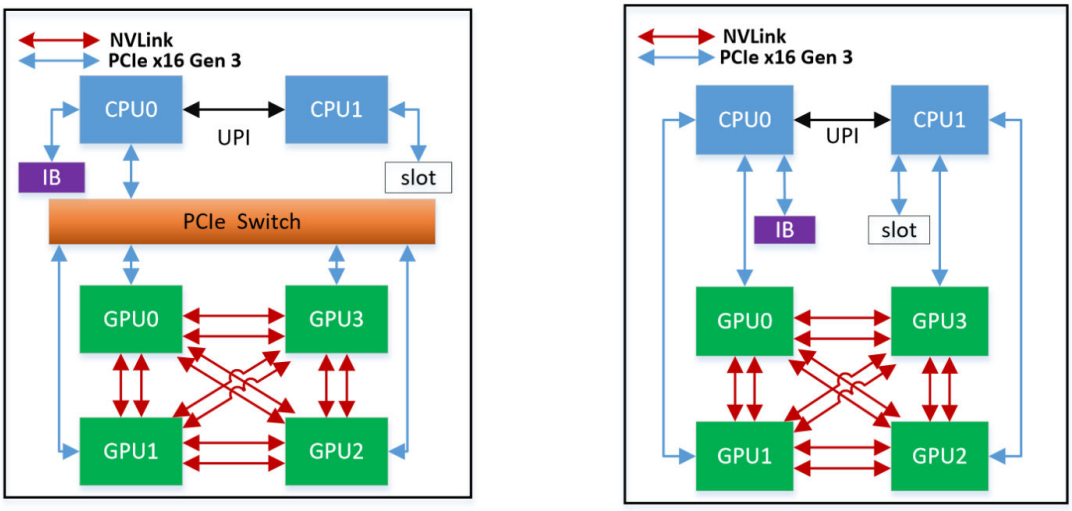
4 Memory solutions

With such large amounts of storage state required, it is not possible to keep the data on the GPU processor. In fact, many high-performance GPU processors have only 1 KB of memory associated with each of the processor cores that can be read fast enough to saturate the floating-point datapath. This means that at each layer of the DNN, you need to save the state to external DRAM, load up the next layer of the network and then reload the data to the system. As a result, the already bandwidth and latency-constrained off-chip memory interface suffers

the additional burden of constantly reloading weights as well as saving and retrieving activations. This significantly slows down the training time and considerably increases power consumption.

Dell EMC Ready Solutions for AI – Deep Learning with NVIDA v1.1 and the corresponding reference architecture guide were released in February 2019. This illustration will quantify the deep learning training performance on this reference architecture using ResNet-50 model. The performance evaluation will be scaled on up to eight nodes.

In August 2018, the initial version 1.0 of Dell EMC Ready Solutions for AI – Deep Learning with NVIDIA was released. In February 2019, this solution was updated to version 1.1. The main difference is that in version 1.1, the CPU and GPU connection topology has been changed from configuration K to configuration M. The comparison of these two different configurations is shown in Figure 1. Unlike configuration K, which has only one PCIe link between two CPUs and four GPUs, the new configuration M has four PCIe links between them, and the memory size of each GPU has changed from 16GB in Ready Solution v1.0 to 32GB in v1.1.



(a) Configuration K (b) Configuration M

Figure 1: The comparison between configuration K and configuration M in C4140 server

5 Using glasses with built-in cameras and inertial sensors

In this section, we propose to use a camera, [2] its task will be to collect visual information based on which a 3D model of the surrounding space will be built, and with the help of inertial sensors (gyroscopes and accelerometers) to specify the location of the visually impaired in relation to the map of this 3D model. Semantic segmentation of visual content (images received from cameras and point clouds received from lidars) should become an important feature of the smartwatch. The device will not only build a 3D model of the surrounding space but also make its semantic assessment, assigning semantic labels to everything that surrounds the visually impaired. Thus, with the help of these devices, a metric-semantic 3D model of the surrounding space will be built, giving descriptions not only with the help of coordinates but also in terms of streets, buildings, premises, and objects.

Another important feature of this model will be the ability to identify and mark not only static objects but also moving ones, such as people, animals, vehicles, and moving interior elements. For moving objects, the trajectories of their movement will be determined.

Such a model will be easy to understand for a person since it is not just information "such and such a distance to an obstacle." The model will not only facilitate orientation in space but will also contribute to more accurate decision-making, for example, the choice of the optimal route of movement. The metric-semantic model will be divided into separate layers, and the visually impaired will receive a description of the surrounding space from different angles, i.e., a complete semantic picture of the world.

For semantic segmentation, neural networks will be used and trained at the initial stage on open data sets. For known objects, computer models can be prepared in advance using CAD systems. Unknown objects will be "drawn" by means of centroids and bounding paths. Accurately detecting an object is not always possible, so the device

will give a probabilistic estimate of its "confidence" that the object is correctly detected.

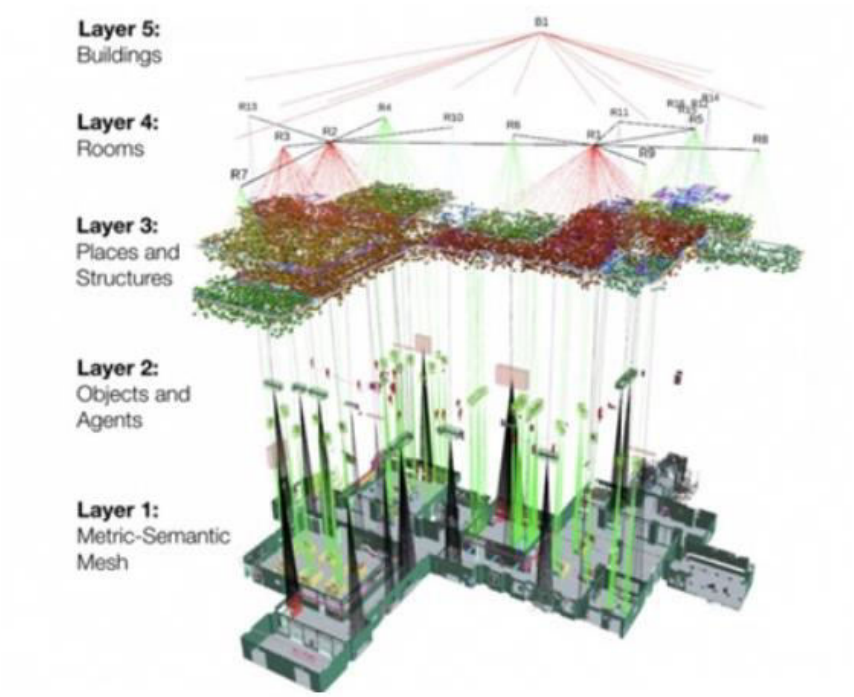
The information collected in the metric-semantic model can be conveyed to the visually impaired both using an interface in the form of an assistant capable of generating audio messages, and through the channels of neuroplasticity, modules to form "alternative vision" using audio or tactile sensory substitution.

6 Next suggested approach is the semantic network

In this section, we will discuss the need to [2] extract data from a neural network and the importance of creating a semantic network of the surrounding space based on them, on which each object will receive the appropriate semantic labels, ceasing to be just a set of color gradients.

If the data is organized into a semantic network, the semantic load is reduced from two objects to one and, therefore, additional "recognition" is not required in the operating mode. The most important thing here is the reduction. Due to this reduction, there is no need to obtain and process a detailed image. Only the main features are fixed.

Imagine that a detailed semantic network describing all possible types of billboards is pulled together into a single node labeled "billboard". The computational complexity in the operational mode will decrease dramatically, which means that when using relatively inexpensive computing resources, the speed of processing and decision-making about control will increase.



Unified representation of spatial perception in the 3D Dynamic Scene Graphs model

7 Results

We are proposing a game changer in the realities of Visually Impaired people. Using our Technology, both elderly and young vision disabled will enjoy a simple way of living and "seeing" without the need of External Factors. The independent factor is extremely important for all human beings and especially when it comes to disability challenges.

We believe in our technology and the benefits of it and assume the rates of success will be high. As our way of discussing how successful our program will be is a test taken by 200 Visually Impaired people who are simulated in daily life challenges. The better they get along with obstacles on the way, the better we are as a technical solution for the blind.

8 Discussion

We believe in making humanity better for all of us and for that we will strive to having our technology as relevant as could be for the disabled

Visually Impaired people. We stand for new technologies and deep learning strategies to impower our technology to renew its data and improve its Images and videos on daily basis. The more the hardware & software will absorb from the surrounding – the more data it will collect and expand its knowledge to the user.

We hope for 20% worldwide users until 2025 and 50% until 2028.

We believe a big change needs to be done in this increasing humanity problem.

References

[1] https://ieeexplore.ieee.org/document/8553750

[2] https://www.iksmedia.ru/articles/5717318-Iskusstvennyj-intellekt-kak-povodyr.html

[3] https://arxiv.org/abs/2005.04473