import.io

Using Web Data to Power Deep Learning Breakthroughs

The Opportunities and How You Can Capitalize

import.io

Executive Summary

Once the sole purview of academics and a few of the largest high-tech companies, deep learning now represents an approach that's poised for rapid and widespread growth across a range of companies and industries. This paper offers an introduction to deep learning, providing an overview of how it has evolved and its key building blocks.

A Brief Introduction to Artificial Intelligence, Machine Learning, and Deep Learning

Artificial Intelligence

Artificial intelligence (AI) was a concept that was introduced in the 1950s. Initially, AI was inherently a rules-based approach. Developers would create rules that would be used to process information. The theory was that if you wrote enough rules, you'd ultimately have an AI system that could work on its own to solve complex problems. However, over the years, the promised benefits of these rules-based theories weren't realized in practice. It was simply too difficult to write enough rules, with the details required, to generate the intended results.

The Entry of Machine Learning and Deep Learning

Over time, others began to take a very different approach to AI, one based on machine learning. With machine learning, the approach was to work from data, and use statistics to extract patterns and relationships from the data. In particular, one subset of machine learning called deep learning has gained prominence. Today, when it comes to AI, virtually all popular attention is being focused on deep learning, which in the span of a few years has fueled a number of breakthrough innovations.

It is deep learning that is now enabling computers to surpass humans in areas like understanding voices and assessing and categorizing images. Deep learning is one of the chief advancements that is powering such innovations as self-driving cars.

It is deep learning that is now enabling computers to surpass humans in areas like understanding voices and assessing and categorizing images.

Deep Learning Defined

Deep learning is the process of running a massive amount of data through a large network of artificial neurons, which are effectively tiny pieces of code. These neurons each hold specific numbers, called weights, which represent how much credence

they give to their neighbor neurons. These weights may initially be random, but as the network absorbs large numbers of examples, these weights begin to be quantified. This is how deep learning implementations can ultimately be trained to execute specific tasks. Once trained, these neurons are no longer modified, and the network can now answer questions on new data, such as classifying images.

There are three different types of deep learning:

- **Supervised learning**. With this approach, large volumes of data are fed into the network. For example, millions of tagged photos can be imported, and supervised learning can then begin to classify the visuals. Most commercial applications use supervised learning.
- **Reinforcement learning**. Using this type of deep learning, the network gets training data by interacting with a system, for example, by playing a game or driving a vehicle.
- **Unsupervised learning**. Instead of being given rules, data is fed in and the network is relied upon to find correlations.

Practical Applications of Deep Learning

Today, there are a virtually unlimited number of ways companies can start to leverage deep learning to further their business objectives. This is true not just for the largest enterprises, but for mid-size and smaller businesses as well. Deep learning has been used to create applications that would have sounded like science-fiction a few years ago. In large part, this is because of breakthroughs in image and speech recognition. Deep learning can also be used to improve a process by reducing the error rate of a model. For example, an organization can employ deep learning, so that, instead of relying on a few simple assumptions, rich intelligence can be used to improve fraud detection.

To illustrate the range of possibilities, following are a few examples:

- **Retail**. An online retailer featured customer reviews of all its offerings. The management team wanted to analyze these reviews, and the reviews of customers on competitor's sites, in order to gauge customer sentiment, and track how it was changing over time. With deep learning, the retailer was able to feed the text and ratings from these reviews into the deep-learning network, thereby creating a high-precision classifier for reviews. They could then track how sentiment varied over time.

- **Manufacturing**. A manufacturer with an online products catalog wanted to identify all the products featured on the site and match them to their in-house taxonomy in order to improve its promotional efforts. The team leveraged deep learning, starting by leveraging data from multiple sites with similar products, including product images, titles, captions, and descriptions. The team could then fine-tune an existing classifier trained on ImageNet, and map captions to their taxonomy.
- **Travel**. A travel site's management team wanted to deliver an application that offered customers insights into anticipated travel times. They kicked off a machine learning project that leveraged the structured data from airlines and airports relating to scheduled and actual departure and arrival times. By leveraging this information, the team could deliver an application that would make recommendations in terms of optimal airports, airlines, and travel times in order to minimize delays.

Ingredients of Deep Learning

If your needs are simple, chances are you can use a standard hosted solution for deep learning. Amazon Web Services (AWS), Google, and a growing number of other companies are offering cloud-based APIs that address common requirements: recognizing objects in an image, detecting faces and emotions, understanding speech and what is being asked, extracting concepts from text, and more. It's as simple as integrating an API into your application—no deep learning expertise needed.

If your requirements are more specific, you need to understand more about deep learning. The good news is that most of the key ingredients are readily available. Here's a list of the core components that make up a deep learning initiative.

Math

In the early days, this would have required math experts with extensive deep learning expertise. Now, you can leverage universal libraries that are easily accessible, so your team doesn't have to understand the math theory to implement deep learning.

Hardware

Deep learning requires high capacity processing capabilities in order to do billions of calculations per second, and to do so repeatedly for a sustained period. Now, it is more practical and cost effective than ever to leverage this kind of processing power. In recent years, the graphics processing units (GPUs) developed initially for gaming have yielded significant advancements in performance. Further, deep learning computing resources can now be obtained as a service from cloud providers like AWS.

Toolkits

Today, deep learning toolkits are available for free from various technology vendors and academic institutions. These kits allow you to build models from scratch or download an existing model. (Do a search on the phrase "model zoo" to see an example.) You can then tinker with the model until it addresses your requirements.

People

In recent years, undertaking a deep learning initiative would require teams of experts with Ph.D.'s in math. Now, all a successful effort takes is programming skills. In addition, there's a very helpful user community, including forums where you can get questions answered, meetups where you can talk to experts, and sites where you can access sample code and models.

Data

Data is the key ingredient to an effective deep learning effort. At a high level, the bigger and more accurate the data sets, the more precise the deep learning results will be. To do an effective supervised learning effort, you might need millions of examples for training your models.

When it comes to sources for data, options run the gamut. Some sources, such as Wikipedia and Common Crawl can provide sources of text that are large and current. DBPedia provides structured data describing millions of things and facts. Pretty much any type of images can be found through search engines and specialized sites. Google has released large data sets around text, images, and videos as well.

Searching for more specialized data sets online will return thousands of options, most of them freely available. However, these data sets tend to be small and can quickly become outdated. For specialized needs, you will need to assemble your own training data, and web content provides a virtually unlimited source of data for deep learning. The advantages of web data are that it's free, often live and current information, and extensive.

However, the downside is that the data available will often need some level of cleanup in order to effectively support deep learning. The exact nature of the cleanup depends on the data, but the idea is that this data is used as examples to teach the neural network.

Let's use an example. Imagine you want to train a classifier to detect fire trucks. First, you need to collect lots of images of fire trucks, as well as other things that are NOT fire trucks, and make sure each image is labeled accordingly. Second, you need to make sure this data set is diverse: collecting lots of images is useful only if they show the

variety of fire trucks, the different shapes, sizes, colors, accessories, and so on. Should your classifier detect plastic toy fire trucks? If so, label them as such. If not, label them as "other."

The same criteria would apply to text data. If you want to train a sentiment analyzer on customer reviews for your site, you need to collect many such reviews with an associated five-star rating. You should get these reviews from sites from the same domain: the language used in hotel reviews may differ from reviews of consumer electronics.

Conclusion

In recent years, tremendous breakthroughs have been realized in the AI domain, specifically in the area of deep learning. Deep learning is fueling many of the most prominent innovations coming to market, including in such areas as driverless cars and image and speech recognition. Where in the past, deep learning would require a massive investment and teams of Ph.D.'s, a number of innovations have put this powerful approach within reach of virtually any organization.

About Import.io

At Import.io, our mission is to deliver intelligent technology that translates the web into data, so you can better understand your world. Hundreds of global businesses rely on our patented technology to get high-quality, high-volume data from the web, without breaking the bank. With Import.io, you can create your own data sets in minutes, without any coding.