УДК 004.451:811.111

Y. V. Opolskyi
L. V. Tulchak

MACHINE LEARNING TOP TRENDS IN 2017

Vinnytsia National Technical University

Анотація

Розглянуто найвагоміші напрямки розвитку сфери машинного навчання та штучного інтелекту у 2017 році.

Ключові слова*: машинне навчання, штучний інтелект, набори даних, тренувальні дані, синтетичні дані, кібербезпека, алгоритми.*

Abstract

The most important areas of development of the field of machine learning and artificial intelligence in 2017 are considered.

Keywords*: machine learning, artificial intelligence, datasets, training data, synthetic data, cybersecurity, algorithms.*

Machine Learning has revolutionized the world of computers by allowing them to learn as they progress forward with large datasets, thus mitigating many previous programming pitfalls and impasses. Machine Learning builds algorithms, which when exposed to high volumes of data, can self-teach and evolve. When this unique technology powers Artificial Intelligence applications, the combination can be powerful. We can soon expect to see smart robots around us doing all our jobs – much quicker, much more accurately, and even improving themselves at every step. Will this world need intelligent humans anymore or self-thinking robots soon outclass us? What are the most visible 2017 Machine Learning trends? [1]

The rise of fake data

According to Crowdflower's 2017 Data Scientist report, 'when asked to identify the biggest bottleneck in successfully completing AI projects, over half the respondents named issues related to training data such as "Getting good quality training data or improving the training dataset", while less than 10% identified the machine learning code as the biggest bottleneck.'

Collecting sufficient data together to build a training set to build AI is easy today, but making sure it is of high enough quality far more difficult. It is also far more important, and data scientists are currently spending the bulk of their time cleaning, labeling, and categorizing data to get it up to the quality they need. One solution to this that we are seeing increasingly touted that is set to become more prominent this year is synthetic data.

Synthetic data is artificially produced data that mimics more or less exactly the properties of real data. There are two primary ways to generate synthetic data. The first is by observing real-world statistic distributions from the original data and reproducing fake data by drawing simple numbers. The second is by creating a physical model that explains the observed behaviour, then reproducing random data using this model. For example, if you applied a generative model built on a neural network to the number of images of faces for the purposes of facial recognition, it would produce fake images of faces. This could be applied to a wide range of other data, establishing patterns and then producing something that fit into the range established. Therefore, real datasets are needed to work and they will never be replaced entirely. No model will ever be able to generate examples of things if it has never seen real ones before.

One early stage startup making ground in the field is Automated DL. The Virginia-based company creates synthetic data by using generative models that create data that resembles or are in some way related to historical examples they are trained on. We expect to see a number of others start to make moves in the area as the year progresses.

Democratization of machine learning becomes more important

The voices to democratize AI have come from leaders from every leading tech companies. Microsoft CEO Satya Nadella recently wrote that he wanted AI 'in the hands of every developer, every organization, every public sector organization around the world' to allow them to build their own intelligence and AI capability. Fei-Fei Li, chief scientist of artificial intelligence and machine learning at Google Cloud, agrees, stating 'The next step for AI must be democratization. This means lowering the barriers of entry, and making it available to the largest possible community of developers, users and enterprises.'

Now, Google has TensorFlow, the open source set of machine learning libraries opened in 2015. Amazon has made its Deep Scalable Sparse Tensor Network Engine (DSSTNE - pronounced 'Destiny') library available on GitHub under the Apache 2.0 license. Elon Musk has OpenAI, which considers itself as a 'non-profit AI research company, discovering

and enacting the path to safe artificial general intelligence'. Google also recently announced its acquisition of online data scientists' community Kaggle, which is a community of data scientists but also one of the largest repositories of datasets that will help train the next generation of machine-learning algorithms.

The cloud will also help in the push for democratization. Machine learning requires an immense amount of computing power to function correctly - power that was previously out of reach for many companies. The scalability offered by Amazon Web Services (AWS), Google Cloud Platform, Microsoft Azure enable developers to build out their infrastructure optimized for machine learning at a fraction of the cost of developing their own proprietary system.

Machine learning becomes vital for cybersecurity

As we saw in the recent WannaCry attacks, cybersecurity poses a clear and present danger to organizations in every industry and it is unlikely to be resolved anytime soon. Research by Accenture found that the average organization faces 106-targeted cyber-attacks per year, with one in three of those attacks resulting in a security breach. Towards the end of 2016, estimates put the number of new malware samples being generated in a single quarter at around 18 million - as many as 200,000 per day.

This threat is constantly mutating, as hackers adapt to cybersecurity measures and find new ways to infect systems. In order to deal with this, organizations must be extremely quick to adapt their security countermeasures, and machine-learning techniques are the only technology currently available with this capability. Former Department of Defense Chief Information Officer, Terry Halvorsen, believes that 'within the next 18-months, AI will become a key factor in helping human analysts make decisions about what to do.' This point of view is being reinforced by significant investment in the field by the world's largest technology companies.' According to DFLabs's May 2017 report 'Next Generation Cybersecurity Analytics and Operations Survey,' 93% of IT leaders are using or planning to use these types of solutions, 12% have deployed machine learning technologies designed for security analytics and operations automation and orchestration, 27% that they're doing so on a limited basis, and 22% said they're adding them. Just 6% said they are either not planning on or not interested in deploying these technologies.

MIT has been experimenting with it for some years, while IBM is training its AI-based Watson in security protocols and has now made it available to customers. Amazon also recently acquired AI-based cyber-security company Harvest.ai, which uses AI-based algorithms to identify the most important documents and intellectual property of a business before combining user behavior analytics with data loss prevention techniques [2].

Evolution Algorithms make a Comeback

For supervised learning, gradient-based approaches using the back-propagation algorithm have been working extremely well. Moreover, that is not likely to change anytime soon. However, in Reinforcement Learning, Evolution Strategies (ES) seem to be making a comeback. Because the data typically is not iid (independent and identically distributed), error signals are sparser, and because there is a need for exploration, algorithms that do not rely on gradients can work quite well. In addition, evolutionary algorithms can scale linearly to thousands of machines enabling extremely fast parallel training. They do not require expensive GPUs, but can be trained on a large number (typically hundreds to thousands) of cheap CPUs.

Earlier in the year, researchers from OpenAI demonstrated that Evolution Strategies could achieve performance comparable to standard Reinforcement Learning algorithms such as Deep Q-Learning. Towards the end of the year, a team from Uber released a blog post and a set of five research papers, further demonstrating the potential of Genetic Algorithms and novelty search. Using an extremely simple Genetic Algorithm, and no gradient information whatsoever, their algorithm learns to play difficult Atari Games.

AI & Medicine

The year 2017 saw many bold claims about Deep Learning techniques solving medical problems and beating human experts. There was a lot of hype, and understanding true breakthroughs is anything but easy for someone not coming from a medical background. Among this year's top news there was a Stanford team releasing details about a Deep learning algorithm that does as well as dermatologists in identifying skin cancer. Another team at Stanford developed a model that can diagnose irregular heart rhythms, also known as arrhythmias, from single-lead ECG signals better than a cardiologist can.

Nevertheless, this year was not without blunders. DeepMind's deal with the NHS was full of "inexcusable" mistakes. The NIH released a chest x-ray dataset to the scientific community but upon closer inspection, it was found that it is not suitable for training diagnostic of AI models.

Self-driving cars

The big players in the self-driving car space are ride-sharing apps Uber and Lyft, Alphabet's Waymo, and Tesla. Uber started out the year with a few setbacks as their self-driving cars missed several red lights in San Francisco due to software error, not human error as had been reported previously. Later on, Uber shared details about its car visualization platform used internally. In December, Uber's self-driving car program hit 2 million miles.

In the meantime, Waymo's self-driving cars got their first real riders in April, and later completely took out the human operators in Phoenix, Arizona. Waymo also published details about their testing and simulation technology.

Lyft announced that it is building its own autonomous driving hard- and software. Its first pilot in Boston is now underway. Tesla Autpilot has not seen much of an update but there is a newcomer to the space: Apple. Tim Cook confirmed that Apple is working on software for self-driving cars, and researchers from Apple published a mapping-related paper on arXiv [3].

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. 2017 Machine Learning Trends [Електронний ресурс] – Режим доступу: http://www.dataversity.net/2017-machine-learning-trends/ - Назва з екрану;

2. What is 2017 bringing for machine learning? [Електронний ресурс] – Режим доступу: https://channels.theinnovationenterprise.com/articles/machine-learning-top-trends-in-2017 - Назва з екрану;

3| AI and Deep Learning in 2017 – A Year in Review [Електронний ресурс] – Режим доступу: http^||www|wildml|com|2017|12|ai-and-deep-learning-in-2017-a-year-in-review|- Назва з екрану.

Опольський Ярослав Віталійович **—** студент групи 2СІ-14б, факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, електронна адреса : opolsky.yarik@gmail.com

Науковий керівник: ***Тульчак Людмила Володимирівна*** **—** *с*тарший викладач англійської та німецької мов, кафедра іноземних мов, Вінницький національний технічний університет, Вінниця

Opolskyi Yaroslav V. **—** Student of group 2SE-14, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, email: opolsky.yarik@gmail.com

Supervisor: ***Tulchak Liudmyla V.*** **—** Senior Teacher of Foreign Languages Department, Vinnytsia National Technical University, Vinnytsia