Current A.I. Trends: Hardware and Software Accelerators

Tech Survey

MATTEO PRESUTTO

Royal Institute of Technology

Minor's Degree Project
Stockholm, Sweden May 2018

Abstract

AI is evolving at a pace quicker than ever in history: private and national founds are being invested to embed it in services while the research community keeps pushing state of the art boundaries.
Hardware and software accelerators are playing a key role in such a dynamic environment where computational advantages can be directly translated into business opportunities by speeding up model deployment and unlocking the potential of big data.

An A.I. accelerator is any hardware or software product made to speed up the computation of A.I. tasks. In this survey, the status quo of current A.I. trends is described with particular focus on the hardware and software accelerators side. The document starts by defining what A.I. is, a survey of A.I. based products and services is then introduced to the reader followed by a survey of hardware and software accelerators. We will see how pains of A.I. based products are well complemented by accelerators, how software and chips are reshaping industries and that cloud computing is playing an important role in today's A.I. market scenario.

The survey closes with a section where future work is discussed, emphasis is put on what has not been answered in the paper and on recapping the previous chapters.

A thank goes to my supervisor Bo Karlsson for the precious feed-backs, to Andreas for supporting me during my M.Sc. thesis and to Maria-Teresa for her tips. Many friends and colleagues contributed indirectly to the content of this document too with suggestions and discussions, I'm grateful to you all for your help.

Contents

1 Introduction

Marvin Minsky defined AI as *"the science of making machines do things that would require intelligence if done by men"*. Even if the field has existed ever since the beginning of Computer Science, the description of what intelligence is is still not well defined and no single meaning has been unanimously accepted by the community. This is reflected in the number of AI sub-fields, which differ philosophically and methodologically and often fail to communicate with each other.

The AI field notoriously stalled in a depressive age during the 80s and 90s, this has been attributed to computational unavailability (computers were not fast enough) and to a consistent failure in matching the high expectations of the industries on such matter. In those decades, research and development has been slowed down by cuts on funding and investments [7] [11] [5].

In the last decade, AI has been able to leverage the higher computational power available to advance research, which in turn proved the field potential with novel techniques like Deep Learning, Search and Information Retrieval.
Major AI investments are now being allocated by the nations: namely the EU has recently called for a €20 billion investment through Horizon 2020, the CAS Institute of Automation in China allocated $ 150 millions for AI R&D and companies in USA raised a volume of $ 12 Billion in 2017 alone.

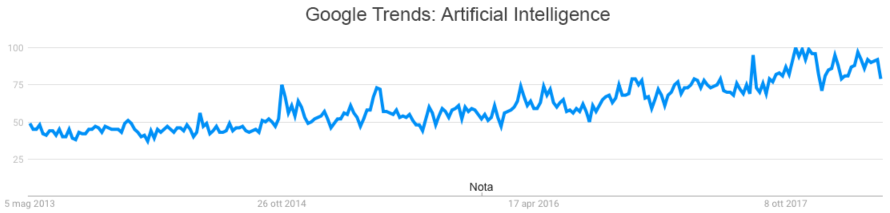


Figure 1: Google trend index for the keywork "Artificial Intelligence" over time

As AI growth is accelerating (Figure 1), its hyping status is having an impact in all sectors and multinationals are competing for AI domination.

1.1 AI Subfields

The fragmented and dynamic nature of AI makes it difficult to identify its subfields. AI domain sub-components, as described by Ming-Hwa [10], comprehend:

- Machine Learning
- Computer Vision
- Data Mining
- Information Retrieval

- Speech Recognition and Natural Language processing
- Robotics
- Search and Optimization
- Knowledge representation
- Logic and Probabilistic reasoning
- Expert Systems

Machine Learning is the science which studies computational learning algorithms. What learning is is well described by Tom Mitchell in his book "Machine Learning": *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P,if its performance at tasks in T, as measured by P, improves with experience E.*

Computer Vision is a field whose objective is to make computer able to process images and videos in a manner equal to the human one. Typical examples include object detection, semantic segmentation and video tracking.

Data Mining is the process of discovering knowledge from data. It is closely related to Machine Learning, Big Data and Statistics.

Information Retrieval (IR) is the process of retrieving information from a set so that it is relevant to the user need. Search engines are a direct application of IR, this makes it closely related to the Search sub-field.

Speech Recognition and Natural Language processing is the sub-field of AI that studies how to interpret speech and natural language in a manner equal to the human one. Applications range from synthetic speech production to speech captioning, language translation and chat-bots.

Robotics is a field of engineering that studies how to construct, operate and design robots. Recent advancements in AI are innovating such field with tools like Reinforcement Learning and Computer Vision.

Search and Optimization is the field of mathematics whose aim is to develop algorithms that efficiently search solutions to mathematical optimization problems.

Knowledge representation is the sub-field of AI that studies how to better represent knowledge in a human readable format.

Logic and Probabilistic reasoning is a sub-field of Mathematics whose objective is to use Probability theory to model uncertainty of logical propositions using deduction.

Expert Systems are decision-making models that emulate human reasoning. They are one of the very first application of AI and have been around for almost 50 years.

1.2 The State of the Art

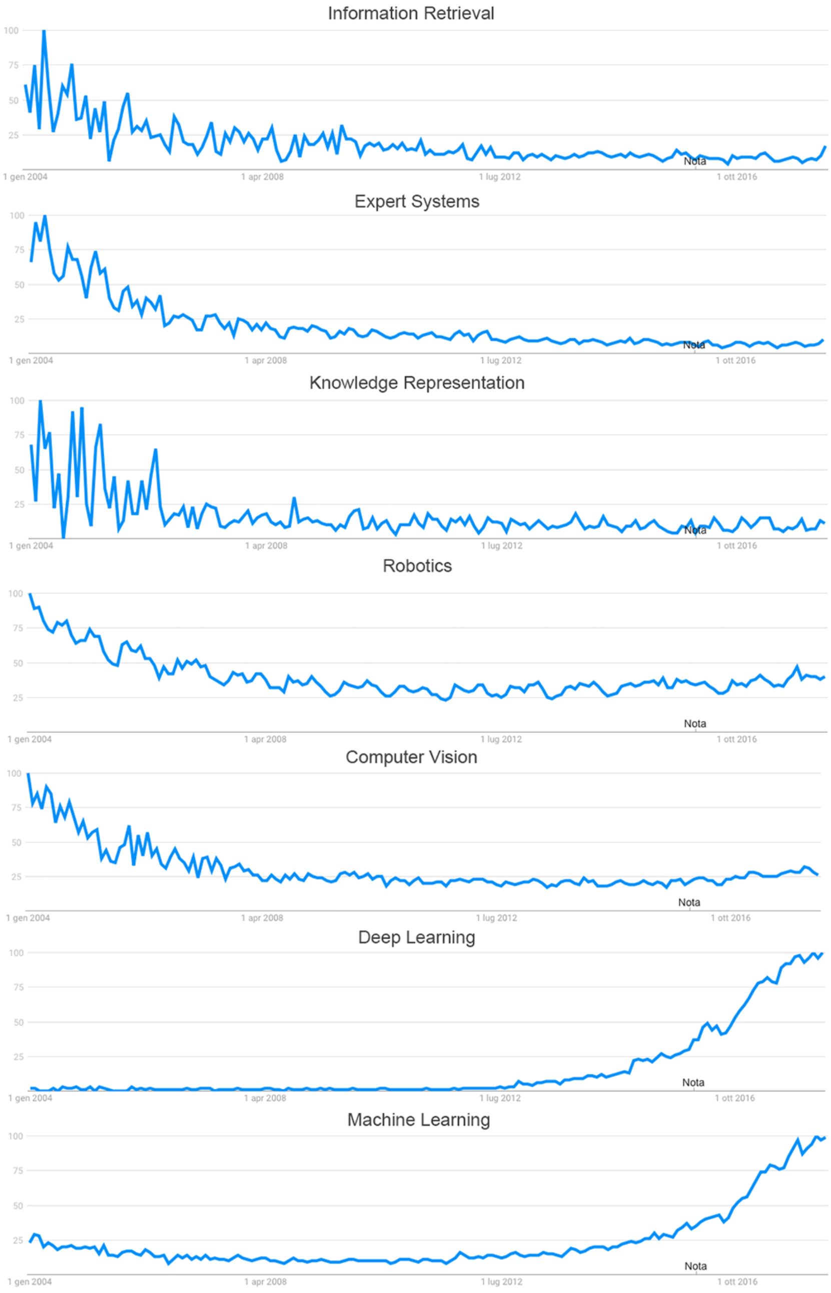
A closer look at the sub-trends in AI shows an interesting pattern: as shown in figure 2, all sub-fields except Deep Learning and Machine Learning experienced substantial decrease in popularity. On the other side, Deep Learning has seen a steep increase closely followed by the Machine Learning trend graph. Since Deep Learning is considered a subset of Machine Learning, it is reasonable to speculate that the hype in Machine Learning itself is a byproduct of the hype in Deep Learning. Furthermore, the modest recent increase in popularity of Computer Vision and Robotics can be explained by the applicability of Deep Learning in those fields.

The Deep Learning trend can be traced back to the influential paper *Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton - ImageNet Classification with Deep Convolutional Neural Networks - 2012* where it was first proven how deep neural networks (i.e. Neural Networks with many layers) could be tuned to substantially outperform classical Computer Vision, feature engineering and Machine Learning algorithms for the task of image classification.

Many important steps ahead have been taken since the publication by A. Krizhevsky et. al. Deep Learning has been shown to be a killer technology in countless fields, to name a few:

- Text processing: processing text. Includes text classification, clustering and summarizing.
- Speech captioning: producing captions from a speech digital audio file.
- Synthetic speech production: producing speech from a text input.
- Image segmentation: classifying pixels in an image.
- Face identification: detecting faces in an image and identifying them.
- Object detection: detecting objects type and location in an image.
- Language translation: translating a text from a language to another.

Finally, it has been proven that in image classification tasks Deep Convolutional Networks significantly outperform human abilities by quality and speed. The dominance of Deep Learning on this particular task set is well established, to the point of being allowed for commercial applications for medical diagnosis [2].


Information Retrieval

Expert Systems

Knowledge Representation

Robotics

Computer Vision

Deep Learning

Machine Learning

Figure 2: Google Trends over time for different AI sub-fields keywords

2 Related Work

Previous work has been done to better understand the prospects of Deep Learning for commercial usage, here we discuss the most up-to-date documents found during the development of this document which might be useful to the reader.

In [9], AI software frameworks are bench-marked to analyze their market potential, the document focuses completely on Deep Learning. The study takes into account several use cases of deep learning, analyzes each framework strengths and weaknesses, their popularity and the hardware supported. The document represent an excellent and up-to-date source of deep learning frameworks, very helpful in the decision making process of choosing the right framework for a given application.

In [1], the market of Deep Learning processors for data center is analyzed in detail with a forecast projection till 2022. It analyses in great detail the players of the market, including startups. Importantly, like other documents cited, this work emphasizes the trend of treating model inference and training as two different problems with different computational needs. The document is a very powerful resource for businesses who need to choose an AI Hardware accelerator to support their product and/or service.

In [3], the outcomes of the Mobile World Congress (MWC) are summarized with emphasis on the impact of AI in the telecommunication sector. It is an up-to-date compact resource for telecommunication companies to understand the level of competition and/or cooperation of AI in this industry.

In [4], a market research of Deep Learning as a whole is documented. The paper describes in high detail all the aspects of the market: this ranges from a PESTEL analysis to the never ending list of market estimates and forecasts (till 2025) divided by region, solution, service, end-use and much more.

[8] is a less technical and more informal document which introduces the reader to several trends in AI for 2018. Topics range from the Chinese rise over AI to voice assistants and cyber-security. [6] is a market forecast analysis of storage services for AI purposes.

Finally, the topic and content of this document have been picked for the high compatibility with the author's M.Sc. Thesis. The objective of the M.Sc. thesis is to study neural network shrinkage algorithms for real time hardware devices.

3 AI Based Products

The domain of AI based products is vast and includes non-digital goods like the recent culinary project of Google in which AI has been used to develop a chip cookie recipe. In this analysis, an AI based product is any digital product that implements AI technology. AI based services are not considered and products that merely use AI in their value chain are excluded.

The question arises spontaneously: why use AI for products in the first place?
Apart from the marketing reasons, AI has the ability of delivering value to almost all products in all sectors [8]. Using AI for a product can deliver qualitative advantage, if the sub-task being automatized is in a domain where AI is the state of the art (e.g. image processing), and speed advantage, if the task being automatized is performed by humans.
Not using AI has disadvantages in the long run as the adoptions of it by a competitor has the potential of disrupting the market thanks to the cut in production costs.

Another major advantage of such technology comes as a byproduct of the deterministic nature of AI itself: every task done by humans cannot be processed with causal determinism in mind as it is unreasonable to assume that every employee will perform a task the same way each and every time it is performed (e.g. in the unfortunate case this document gets lost, it would be unreasonable to assume that the author will be able to produce the exact same text even if the topic is the same). With AI it becomes possible to perform the same task the same way each and every time, effectively unlocking the possibility of standardizing the task.

Integrating AI into a product is an iterative an gradual process and details of its implementation are product dependent. Although every solution has its own development course, every AI integration is characterized by at least the following fundamental steps (in no particular order):

- Data gathering: collecting data for decision making and modelling purposes.
- Modelling: the process of creating an AI tool.
- Integration: the process of adjusting the product to integrate the new AI module.

3.1 Products leveraging AI

The amount of products using AI to which we come in contact every day is surprising by itself: from smart-phones to cars, search engines and typing assistants (auto-completion), AI is slowly permeating our lives. More formal examples of what an AI product is are:

- reCAPTCHA is a service for bot detection. It is a binary classifier that challenges the user with simple riddles, it then analyzes the solution provided and the user behavior to output a number describing the likelihood of the user being a human (the higher, the most probable it is a human). The class of riddles it uses is carefully selected by the developers and the problems posed are usually instances of tasks known to be performed better by humans then machines.

- Gmail recently introduced autocomplete tools. This features helps the user write an email by auto-completing sentences on-the-go. As the user begins writing a sentence, the AI will elaborate in background the most probable way of finishing the phrase, the user can then accept the suggestion by pressing ¡tab¿.

- Quetext.com is a webbased plagiarism checking toolkit. It uses AI to check if a text contains plagiarism by searching for matches in billions of documents.

- Dragon NaturallySpeaking is a tool used to input text with speech. Targeting at its core users who suffer from vision loss, this utility can do online captioning from vocal input and read the text.

- CloudFlare is one of many web hosting providers on the internet. One of the cores of its value proposition is cyber security. Distributed Denial of Service attacks are a powerful set of offence techniques used by hackers to disrupt an internet service by overloading it with more requests than it can handle. Cloudfare offers a service of DDoS prevention leveraging AI to detect novelties in the distribution of networking packets and stop the attack as soon as possible avoiding any damage.

- FaceId is a feature recently introduced by Apple for authentication on IPhone. It works by building a face profile of the owner which is used every time the phone is unlocked to check in real time if the user face is matching the owner one. It is able to adjust itself to the gradual changes of the owners face with time.

- SIRI is one of the many voice assistants available today. Its AI capabilities range from voice captioning to chatting. It is able to interact with other application in the Operating System and with the web.

- Amazon product recommendation is a feature of the Amazon platform which suggests users other products based on their preferences and shopping history. It suggests content based on similarity of tastes with other clusters of users and takes into account the similarity of other items with the items shopped.

- Facebook face detection and tagging is a service of Facebook which automatically detects faces in images and tags people from the user friendship set.

3.2 Market Sectors

Simply put, every industrial sector is using AI. The way each sector is using it differs substantially and deserves mentioning.
The very first sectors that started using AI are, unsurprisingly, in the IT domain itself: this includes Search Engines, visualization tool-kits and database management systems. But how are other sectors embracing this technology?

A substantial amount of research efforts has been put in investigating AI in the health-care field. As of today, an extensive amount of medical applications have been developed, products that tackle more sensitive domains are on their way to be validated and commercialized. The health-care field is one of the hottest trends in AI.

We are now starting to see AI being applied massively in the Entertainment industry. It has been applied in gaming ever since its very beginning (notoriously in chess and poker) but many tools have been developed in the medias (e.g. Netflix) and music industry recently (e.g. IBM Watson)

In the agriculture field, AI is being applied to automate production. From automatic quality checks to mineral delivery optimization in hydroponics, the impact of AI in this field its still in its infancy.

More human interactive sectors like Law and Real Estate are slowly adopting the technology too. The state of the art allows for automation of lower level tasks like document summarizing, sentiment analysis and simple decision making sub-processes.

3.3 Pains

As of today, making an AI product is not straightforward and big techs as well as startups have to go through a long process of trial and error before deployment, when deployment is actually plausible.
The pains mostly associated with this technology are:

- Difficulty in using it for Real Time operations due to the computational complexity of AI models.
- The data acquisition process is expensive in terms of either time or budget.
- Training a model requires substantial amount of time, which in turn slows down development.

4 On AI Acceleration

As analyzed in Section 1.2, today's AI hype is a direct product of the rise of Deep Learning (DL) and there is demand for AI speedup services and products. Big tech companies and startups are now emerging to take over this market.

From a more formal point of view, AI accelerators are the set of products whose aim is to speed up AI computation. They are divided in Software and Hardware solutions, usually working together to deliver more value to the customer. Given the hyping status of DL, almost all of the products being developed are aiming at speeding up specifically DL ignoring for the most part all of the other AI algorithms. For this reason, in the rest of this document whenever we cite AI we mostly refer specifically to Deep Learning.

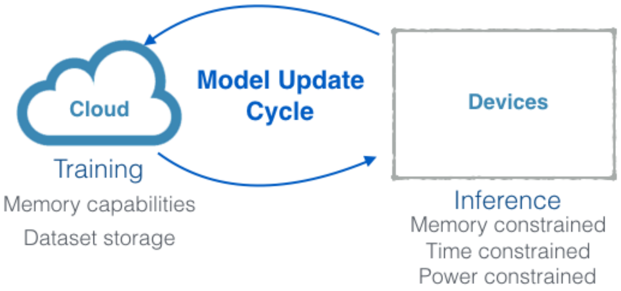
Hardware accelerators are a set of chips designed to run AI algorithms faster than current general purpose solutions. This type of products started to appear in the market at the beginning of the 2000s when GPUs were discovered to be an efficient multi-core alternative to the much more expensive many core CPU systems. The product scene today is starting to spread and converge to different solutions often very different from each other.

The software used in AI is now competing disruptively in speed and capabilities. Different developer communities are doing their best to catch up with the latest trends in academic research by implementing novel algorithms as soon as they are published. At the same time, developers are optimizing the software computational capabilities to as many devices as possible pushing their speed to the limit.

One of the most important recent trends regarding both AI software and hardware is to consider inference and training as two separate different phases of creating a model which require different computational approaches.
To better understand this process, we are now going to define what training and inference are by showing a particular example.
Lets say we have access to a data set of images, half contains dogs and half cats. To train a model to tell if in an image there is a cat or a dog, we feed images to it and adjust its parameters so to output the right answer (either cat or dog depending on the particular image). This procedure is practically and theoretically unpredictable and we can't know a-priori how much time it will take to train the model.
The inference process of an AI model is the procedure of actually using it. In the cat/dog scenario, inference would be the act of labelling images whose content is unknown. This is done by feeding unlabelled images to the model which in turn outputs its predicted class. This process, contrary to the training one, is extremely predictable and it is possible to calculate very precisely how much time it will take to process a given set of images.

In the market, it is becoming the standard to develop specific chips for training and specific chips for inference. Training requires large amounts of data, is time expensive, is unpredictable and requires a certain level of human assistance

and testing. Inference on the other hand doesn't require human assistance nor testing, is predictable but can still be time expensive.
This is why we are seeing the emergence of cloud based services for AI training and fast/small chips for inference. Furthermore, inference is required on device for latency minimization purposes, as is the case for autonomous car decision making or offline face detection on a phone. The AI process chain is well summarized in the following image:



On cloud there is prevalence of wide memory, high power consumption, expensive and powerful computational chips whose aim is to create or improve models while on devices there are chips specialized to run the trained models in a memory, time and power constrained environment like phones, cars or drones. Whilst the train-inference split is consistent among competitors, the way their products work can be substantially different. In the following section we will see a set of different approaches.

5 Hardware Accelerators

An AI hardware accelerator is a physical chip whose aim is to speed up AI computing. The difficulty behind creating such chips lies in the dynamism of AI itself, what algorithms should be accelerated and what not? Answering this question is not a trivial task since building chips requires time and high investment volume and it is very difficult to introduce modifications once chips are deployed (impossible in many cases).

Important for both training and inference is the parallelism: since the amount of computational power required is substantial, it is important to be able to split problems into sub tasks to be spread among different computational units. This process is done inside chips at the hardware level too thanks to schedulers and multi-core processors. Training requires a higher level of communication among different cores than inference due to the fact that gradients need to be averaged to optimize the model. Inference requires less communication among cores but splitting a model still introduces memory overhead since the results of the different cores must be aggregated anyway (more on this point in the software section).

5.1 Hardware Products

The major players in the hardware product scene identified for this document are:

COMPANY	PRODUCTS	MAJOR PARTNERS
Nvidia	DGX series AI GPU series HGX server framework Nvidia Drive Nvidia Jetson	Arm, Amazon, Google, NXP, Baidu, Tesla, Nokia, Huawei, IBM, Microsoft, Alibaba, Tencent
Arm	ML Processor OD Processor Cortex Series CPU	NXP, Xilinx, Google, Intel, Nokia, Huawei, Qualcomm
NXP	i.MX processor	Arm, Nvidia, Xilinx, Google, Baidu, Amazon, Apple, Huawei, Qualcomm
Xilinx	Zinq MPSoCs Ultrascale+	Arm, Baidu, Amazon, Apple, Nokia, Huawei, AMD, IBM, Microsoft, Alibaba, Qualcomm, Mediatek, NXP, Tencent
Google	TPU	Nvidia, Arm, NXP, Intel, Xilinx, Qualcomm, Mediatek, Huawei
Baidu	XPU	Intel, Alibaba, Tencent, Huawei
Tesla	AI Chip (Name TBD)	AMD, Intel
Intel	Nervana Loihi Myriad EyeQ	Baidu, Arm, Amazon, Apple, Nokia, AMD, IBM
Amazon	Echo (custom AI chip)	Nvidia, NXP, Xilinx, Intel
Apple	A11 processor	Arm
Nokia	Reefshark	Xilinx, Arm, Nvidia
Huawei	Kirin 970	Xilinx, NXP, Arm, Nvidia
AMD	Radeon Instinct MI25	Xilinx, Tesla, Intel, IBM, Alibaba, Qualcomm, Mediatek
IBM	TrueNorth Power9	AMD, Intel, Xilinx, Nvidia
Alibaba	Ali-NPU	AMD, Baidu, Xilinx, Nvidia
Qualcomm	Qualcomm AI Engine	AMD, Xilinx, NXP, Arm
Mediatek	APU	Xilinx, AMD

Table: Market Players Analysis

As visible in the table, dozens of chips are being developed. The companies considered here are only a subset of all the big techs that are investing in AI hardware production but many other smaller companies are trying to gain importance too. Among all the chips considered, we can distinguish 4 different types of computing devices: GPUs, FPGAs, Neuromorphic chips and ASICs. GPUs are a family of computing cards with multiple processors and an on board RAM-alike memory support. Historically, they have been optimized and deployed for graphical processing (G.P.U. = Graphical Processing Unit) but have been found suitable for AI purposes and are now massively adopted. FPGAs (Field-Programmable Gate Array) are a class of computing devices whose physical gates are re-programmable. Neuromorphic chips are hardware inspired by neural networks, they implement models to the circuit level of the chip. Finally, ASICs is the family of processors designed to specifically carry out certain algorithms (for Deep Learning, they usually include multiply and accumulate operations and convolutions).

Taking the from another perspective, it is possible to distinguish among solutions that are being developed for cloud training and for edge inference on device. Based on those factors, the products have been classified in the following table:

PRODUCT	CLOUD OR EDGE	CHIP TYPE
Nvidia - DGX series	CLOUD	GPU
Nvidia - AI GPU series	CLOUD	GPU
Nvidia - HGX server framework	CLOUD	GPU
Nvidia - Drive	EDGE	GPU
Nvidia - Jetson	EDGE	GPU
Arm - ML Processor	EDGE	CPU
Arm - OD Processor	EDGE	CPU
Arm - Cortex Series CPU	EDGE	CPU
NXP - i.MX processor	EDGE	CPU
Xilinx - Zinq	EDGE	Hybrid CPU/FPGA
Xilinx - Virtex	CLOUD	FPGA
Google - TPU	CLOUD	ASIC
Baidu - XPU	CLOUD	FPGA
Tesla - AI Chip (Name TBD)	EDGE	unknown
Intel - Nervana	CLOUD	CPU
Intel - Loihi	CLOUD	NEUROMORPHIC
Intel - Myriad	EDGE	CPU
Intel - EyeQ	EDGE	CPU
Amazon - Echo (custom AI chip)	EDGE	Unknown
Apple - A11 processor	EDGE	CPU
Nokia - Reefshark	EDGE	CPU
Huawei - Kirin 970	EDGE	CPU
AMD - Radeon Instinct MI25	CLOUD	GPU
IBM - TrueNorth	CLOUD	NEUROMORPHIC
IBM - Power9	CLOUD	CPU
Alibaba - Ali-NPU	CLOUD	unknown
Qualcomm AI Engine	EDGE	CPU
Mediatek - APU	EDGE	CPU

Table: Product analysis

5.2 About Cloud Computing

An important factor in the evolution of the product scenario of AI Hardware is the presence of Cloud Computing services. Cloud Computing is a set of services accessible through the internet that provides users computing power on-demand.

The Cloud Computing business model works at its best in AI because of the costs associated with acquiring an AI product. For reference purposes, the Nvidia DGX-2 platform costs around \$400.000 for a single workstation, a price not only inaccessible to most people but even to most startups.
With the advent of Cloud Computing it is possible to use expensive AI hardware paying only per usage, this introduces vast possibilities in the scalability of operations: if one more server is needed, all its required to do is just to book another system on the Cloud Computing platform.

But the power of Cloud Computing in the AI setting doesn't only come from the cost cut and scalability, with the help of cutting edge virtualization utilities it is now possible to book a customized hardware setting that better fits the computational needs of the user. Virtualization is a software utility between the OS and the Hardware whose aim is to "virtualize" the hardware so that resources can be paired to satisfy arbitrary user needs.

As we have seen, the model update cycle of an AI model is performed by sending back and forth information from the Cloud to the device. An advantage of using Cloud Computing for training AI models is that it is possible to smoothly distribute servers in targeted countries, speed up the update cycle and favor communication.

6 Software Accelerators

An AI Software Accelerator is any AI software whose aim is to speed up the computation of AI. As of today, the quantity and heterogeneity of those products is substantial: software differs in programming language, targeted hardware, algorithms, abstraction level and objective.

In the next sections real world examples of software accelerators are provided, with particular attention put on squeezing algorithms.

6.1 Software Products

The following table summarizes the most used AI software products.

COMPANY	PRODUCT	MAJOR PARTNERS
Google	Tensorflow	AirBnb, AMD, Nvidia, Uber, SAP, DeepMind, Dropbox, ebay, Intel, Qualcomm, Twitter, Mediatek Arm
Non-profit	Pytorch	Facebook, Twitter, Nvidia, Salesforce, Paristech, CMU, Uber, Stanford, Oxford, NYU, Inria, ENS, VisionLab, Digital Reasoning
Non-profit	Theano	NOT ACTIVE
Non-profit	Torch	Facebook, Google, DeepMind, Twitter, Nvidia, Yandex
Apache	Mxnet	Amazon, Baidu, Intel Nvidia, Dato, Microsoft MIT, Wolfram
Skymind	DeepLearning4J	Skymind, Redhat, Google
Non-profit H20.ai	Scikit Learn H2O	Inria, NYU, Telecom ParisTech, ADP, CapitalOne, Progressive, Comcast, Macys, Cisco
Facebook	Caffe	NOT REPORTED
Microsoft	CNTK	NOT REPORTED
Apache	MLlib	NOT REPORTED
Nervana	Neon	Intel
Fluxicon	Disco	NOT REPORTED
Non-profit	Prom	NOT REPORTED
Non-profit	LIONoso	NOT REPORTED
Numenta	NuPIC	NOT REPORTED
Apache	PredictionIO	NOT REPORTED
Non-profit	OpenNN	OpenNP, Nvidia

COMPANY	PRODUCT	MAJOR PARTNERS
Rapidminer	RapidMiner	Tableau, Redhat
Non-profit	WEKA	MathWorks, NIST, CERN
Non-profit	Chainer	IBM, Intel, Microsoft, Nvidia
Non-profit	Keras	NOT REPORTED
Non-profit	Lasagne	NOT REPORTED
Nvidia	CUDA	Amazon, Facebook, Google, IBM, Los Alamos, Microsoft, Pixar, P&G
Nvidia	TensorRT	SAP, Twitter
Arm	ArmNN	AiOTA Labs, Codeplay, Enigma Pattern, PILOT.AI, Sensory

Table: Product analysis

The aim of those projects is to provide easy access to AI algorithms and accelerate them. The different software packages value propositions can be classified in the following categories:

- Software accelerators targeting specific hardware for computational parallelization purposes.
- Frameworks to simplify the usage of low level API with either module-like wrappers or new programming languages.
- Software to squeeze models and accelerate them.
- General Purpose machine learning and/or data mining utilities.
- Streaming and real-time oriented data mining software.

6.2 The squeezing hype

One of the hottest trends in AI acceleration is to squeeze Deep Learning models to reduce their memory footprint and accelerate them. As of today, most of the academic research being done in this field focuses on squeezing trained models for inference acceleration targeting edge computing.

The reason why squeezing algorithms have not been applied to the acceleration of training procedures regards the complexity of optimization:
Deep Learning models are optimized with gradient descent. As the name suggests, gradient descent requires the computation of the gradient which is known to be more sensible to distortions (e.g. quantization) than the other components of a model.

Squeezing software supports the following algorithms:

- Quantization: Converting floating point parameters to representations with less bits.
- Distillation: Teaching a smaller model the knowledge of a bigger one.

- Pruning: is the procedure of removing parameters from a model.
- Factorization: as neural network parameters are represented by sets of matrices, the factorization procedure consists in synthesizing those matrices by dropping eigenvalues in the factorized representation of those matrices.

Abstracting from the technicality of those algorithms, squeezing utilities are supported in:

- Tensorflow, with the Tensorflow Tiny library which specifically targets edge systems leveraging squeezing algorithms
- TensorRT, developed by Nvidia is an inference engine. Among its utilities there is the quantization to 8 bits.
- Ristretto, a contrib library built on caffe whose purpose is soley to quantize networks.
- Mxnet, gradient compression utility quantizes the gradient to speed up training.
- CNTK netopt, a contrib utility to quantize and factorize models.

The research community is putting a considerable amount of attention in finding feasible and reliable algorithms to squeeze models. The entire submarket is still in its infancy as it is very reliant on the hardware solutions involved which, as we have seen, are still too heterogeneous and belong to a fragmented market.

Although the heterogeneity of the software solutions is still high, a few aspects of this sub-type of algorithms are starting to get recognized as the standard. This is for example the case for the fact that squeezing algorithms require data, that they fall in the category of algorithms that should be run on cloud and that it is important to take in consideration all the metrics losses when parameters are dropped to avoid unwanted behaviors.

7 Conclusions and Future Work

We have seen what AI is, have analyzed the state of the art and explained what has been done so far to better understand the market and solutions being developed.
We have seen a set of AI based products to better understand how they look like, what and where are they used for. We have gone through an analysis of the pains associated with implementing AI solutions.

Based on this, the AI acceleration latest trends, products and solutions have been analyzed both in depth and in breadth to validate the arguments with real world examples. Many hardware accelerators have been identified to understand how they are used for and what are they good at. Software solutions have been collected and classified to analyze their development and value proposition.

Other aspects of the market need to be analyzed to better understand development in AI acceleration. Relevant questions that have not been answered in this document are: what are the applications of each different AI edge hardware? What approaches are being investigated in neuromorphic computing? How do Edge and Cloud communicate and what solutions are being developed for that market? What are the pains of AI accelerators?

References

[1] *Asia-Pacific Deep Learning Processors for Data Center Market, Forecast to 2022*. Frost&Sullivan. Mar. 2018.

[2] Office of the Commissioner. *Press Announcements - FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems*. en. WebContent. URL: `https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm` (visited on 05/11/2018).

[3] Rosalind Craven, James Eibisch, and Francisco Jeronimo. *AI on Show at Mobile World Congress (MWC) 2018*. IDC. Apr. 2018.

[4] *Deep Learning Market Analysis And Segment Forecast To 2025*. Grand View Research. 2017.

[5] Guy-Warwick Evans. "Artificial Intelligence: Where We Came From, Where We Are Now, and Where We Are Going". In: *University of Victoria* (2013).

[6] Ritu Jyoti and Natalya Yezhkova. *Worldwide Storage for Cognitive, AI Workloads Forecast, 2018â"2022.pdf*. IDC. Apr. 2018.

[7] Nils J. Nilsson. "The Quest for Artificial Intelligence: A History of Ideas and Achievements - pg. 408". In: *Stanford University* 29.4 (2010).

[8] *Top AI Trends to Watch in 2018*. CB Insights. 2018.

[9] Jack Vernon. *Benchmarking AI Frameworks - Productization, Market Rationalization, and Ecosystem Development*. ABI Research. Apr. 2018.

[10] Ming-Hwa Wang. *Artificial Intelligence and Subfields*. Santa Clara University, Department of Computer Engineering, Feb. 2017.

[11] Gary Yang. "The History of Artificial Intelligence - AI Winter and its lessons". In: *University of Washington* (Dec. 2006).