Is the application of AI in consumer technology a good decision?

Introduction

Artificial Intelligence or AI is, in very basic terms, the ability of a computer program or machine to think and learn.

It has also been described as a "systems ability to correctly interpret external data and use those learnings to achieve specific goals and tasks through flexible adaptation". This more detailed definition shows the core principles of AI by referencing its adaptation to show how it is a very flexible technology and the wording of "specific goals and tasks" shows how Artificial Intelligence is supposed to be integrated into a system in such a way that it isn't having to guess about the desired outcome.

In the current technology market, Artificial Intelligence seems to be being pushed to the forefront of the advancement of modern technology. On the surface, the extraordinary ability of AI in handling and processing large amounts of incoming data in a short space of time in order to come to a decision, makes a lot of sense when thinking about where to apply this in everyday workloads. On the other hand, in order for AI to function at its peak efficiency, there has to be that large amount of data available to process in the first place otherwise the technology is not being applied in the most optimum way. The challenge that companies attempting to use AI are facing is trying to find the right balance between applying AI in consumer devices and not wasting resources on developing an unnecessarily complicated way of solving basic problems. Now that the market has decided to begin adding AI to as many devices as possible, ranging from fridges and light fixtures to robotic vacuum cleaners and smartphones, it begs the question, "Is the introduction of Artificial Intelligence into every consumer electronic device actually beneficial to the way the devices function?" or is it just another marketing buzzword and gimmick to help designers and manufactures sell as many of their gadgets and gizmos as they can to try and turn the largest profit?

Hardware enabled AI devices

Huawei Mate 10 Pro

One more affordable consumer device that utilises hardware based AI is the Huawei Mate 10 Pro smartphone.

To power the AI, it uses a uses a Neural Processing Unit (NPU) built into its chipset, the Kirin 970, which also contains the CPU (Central Processing Unit) and IGP/GPU (Integrated Graphics Processor/Graphics Processing Unit). An NPU is a type of processor designed with artificial intelligence in mind which means it is much more powerful in AI tasks than the CPU's and GPU's that were available at the time of launch. The main marketing point for the AI in this phone is the smart camera adjustments based on what the sensor is being pointed at. The phone uses the NPU to adjust settings in the camera app by analysing the image to correct exposure and focus in order to improve the photo being taken. By the means of AI powered object recognition, the NPU has the ability to easily recognise the subject of the photo and make adjustments in order to make the photo look better. To give an example, if the phone recognises food being in front of the camera, it will begin to up the contrast and saturation to a certain point which is calculated by the AI, to make the dish look more appealing and if animals are recognised, the AI knows that it should focus on the fine details of the creature which will make the shot crisper and appear more true to life. The NPU can use the AI powered object recognition to identify the subject of a portrait style photo and apply a bokeh depth of field effect to the image, which keeps the person in focus and makes the background blurry. By using the AI, the phone can also do a lot of the processing after the image has been taken by retro actively identifying the edges of the subject and blurring the outside.

The second main use that is targeted towards the consumer is the use of the NPU to manage resources more efficiently. This means it can keep the phone cool whilst charging to protect the battery and give it a longer life. By managing the resources properly, it can also prioritise foreground processes which will make the phone feel snappier to the user. The NPU is also utilised in translation because of its power when analysing large amounts of data to gather more from context when there are multiple translations of a word so that a more understandable translation is produced.

The main advantage of AI in this smartphone is improving the user experience. Being able to automatically adjust the camera to take better photos is a huge selling point and is one that the user will be able to appreciate when their photos come out sharper and better looking than with the use of AI. The other main advantage is the perceived increase in performance from the better resource management. Whether it's the increased responsiveness from dynamically preventing background processes from hogging the system resources and slowing down the foreground processes or whether it's the battery maintaining its peak capacity for longer so the user doesn't notice a reduction in on battery usage time over the life span of the phone, the the more average user is going to notice these improvements without having to have much knowledge of how the device works.

The main disadvantage is the price. At launch, it cost £699 which is quite expensive considering it only has 6GB of RAM and 128GB of internal storage. Many similar

brands like OnePlus and other chinese phone manufacturers are offering devices with higher specs including twice the internal storage for over £100 less. This means that the NPU aspect of the chipset is either very expensive to design and manufacture or that the addition of the NPU is being used as an excuse to bump the price and increase the profit margins from the phones sales.

Nvidia Turing Graphics Cards

The another implementation of artificial intelligence hardware in a mainstream consumer product is the addition of Tensor cores as part of the latest graphics processing unit (GPU) architecture, code named "Turing" after the famous British computer scientist and crypto-analyst Alan Turing, from Nvidia, who have the largest market share for discrete graphics. Tensor cores are processing cores designed for mixed-precision floating-point calculations that have been optimised for deep learning and artificial intelligence uses. As this is the first time a GPU has been designed with specialized hardware for AI, the number of Tensor cores is contributes to a small part of the total processing power of the GPU but as it is just an extra addition rather than the main reason for buying the card, it will not make too much difference to the consumer. With the new addition of the new Tensor cores as well as some other new technologies that are being sold to the public for the first time, the price for these new generation of GPUs is much higher than previous ones but this is quite normal and is considered a sort of early adopter tax for these new technologies.

The main use for the Tensor cores in the Turing architecture is to power an AI dedicated to applying a new type of anti aliasing, called Deep Learning Super Sampling (DLSS), to the graphics in video games. Anti aliasing is a form of post processing designed to smooth out jagged edges from horizontal lines. DLSS uses an AI to inspect the image to identify the edges and then focus the anti aliasing processing on those areas which can reduce the load on the main GPU or improve the quality of the rendered graphics by reassigning the no-longer used main GPU cores back to the main rendering. Nvidia uses very powerful supercomputers in their server farms to inspect hours and hours of gameplay, and, by comparing the differences between un aliased and perfectly aliased frames, can teach an AI how to recognise areas of the frame to focus on to improve the effect of DLSS. The trained AI's are released to the public through driver updates so that the DLSS technique is always improving for supported games. Another use for the Tensor cores is for rapid development and training of an AI that is being made by someone who doesn't have access to the same level of servers that large companies with large budgets do. By integrating the Tensor cores onto a graphics card means that a high power workstation can have access to this technology without spending extra on a dedicated Tensor core expansion card.

The main advantage of this addition to a consumer graphics card is the better allocation of resources. The graphics card can now use more of its traditional 3D compute cores for the rendering of the main graphics and can offload the anti aliasing to the Tensor cores. As some of the main 3D graphics cores are no longer being used for the post processing effects like anti aliasing, the main graphics rendering has more resources to be powered by than if it also had to do the post processing effects therefore, it can run at higher resolutions, frame rates or graphics qualities without having to turn off any other effects. Another advantage is that the additional Tensor cores allow the graphics card to be better utilised in the development and creation of new artificial intelligences by smaller developers.

One drawback of the way Nvidia is distributing DLSS is that it only works on a game by game basis meaning that if a game has not been analysed in Nvidia's supercomputer, the DLSS is unlikely to work and, if it does work, the effect is going to be less noticeable to the user. However, this drawback is more about the delivery of the trained AI rather than a criticism of putting the dedicated hardware on the card.

Software/cloud based AI

Some artificial intelligences are not powered by the device they appear to be running on and are instead usually based in "the cloud" or on a server in a large data center somewhere else in the world and the client device sends commands to the servers AI so that it can respond.

Common applications of these cloud AI's are in low power voice interactive speakers such as the Amazon Echo line up and Google Home products. These smart speakers have basic commands for various activities including:

- music playback,
- making to do lists,
- setting alarms,
- streaming podcasts,
- playing audiobooks,
- providing weather, traffic, sports and news
- searching the web for information and reading out the results

Another place a cloud based AI is likely to be found is on a smartphone in the form of a "virtual assistant". These can usually do the same things as a smart speaker with many more commands available with the addition of a screen and a sim card. These additions include:

- opening applications
- playing movies and tv shows through popular media streaming sites
- plan routes with directions for navigation
- sending texts and starting calls.

One of the features offered by Googles Virtual Assistant is an ability to make phone calls on behalf of the user named Google Duplex. Currently, the main applications of an automated assistant that can make phone calls are to book appointments at places like the doctors office, to place a reservations at a restaurants, or the book a room at a hotel. The assistant uses AI for vocal recognition to work out what the person on the other end said and to say a good response based on the last thing that was said. Other thing the AI is utilised for in the process of making the automated calls is speech fluency. Duplex introduces "speech disfluencies" into its computer generated speech to make the voice sound less robotic and to help the conversation flow with more fluidity. "Speech disfluencies" are small vocal breaks in the flow of an individual sentence. These vocal breaks are usually a type of non-verbal, vocal sounds such as "um" and "err" to cover when the brain takes a bit longer to formulate a word and sounds like "mm-hmm" to show acknowledgement of the last sentence when a suitable word or phrase isn't available.

One advantage of using a cloud based AI service is that the provider can have as much processing power dedicated to the AI as they need since they are not restricted by the space limitations of a modern smartphone or smart speaker. This means that the AI can become more and more complex based on the needs of the provider and their users whilst still being accessible to those who are not using the latest and most powerful client devices as the system requirements to run the user interface are often quite low.

One disadvantage of running the AI on a separate device is that the providers of the AI services and deliver parts or all of the functionality as a monthly subscription despite the AI being advertised as being "built in" to the client device.

comparisons

costs of hardware vs cloud-Based AI

One of the main comparison points between hardware enabled Artificial Intelligence and Cloud based Artificial Intelligence is that of cost. With hardware enabled AI, the main cost is usually a one time purchase of the hardware device that has the processing cores for the AI. On one hand, the one off cost of the hardware is good since it guarantees the availability of the AI processing benefits whilst the hardware is functional. On the other hand, the one time purchase is of hardware that will, almost certainly, become outdated with new products being launched, with typically

generational improvements to the hardware, and eventually, the hardware that was purchased will need to be replaced to keep up with the power the AI requires for large amounts of data to be processed or the speed requirements that the user or users clients have for the training of the AI. As the hardware will have to be regularly replaced, the large payments at the time of each purchase of a piece of hardware to power the AI will add up over time to a much bigger sum.

With cloud-based AI, the client devices are usually of lower cost, such as a smart speaker which costs in the range of £30-100, but the usability of the AI is more locked down with the voice commands having to be pre-programmed by the providers of the cloud infrastructure that runs the AI. Some providers, such as Amazon with their Alexa AI, let more advanced users develop their own commands and processes that could be accessed from the voice interface but some providers don't let users do this and if they do, there is often a cost associated with hosting your own commands on their servers meaning that if any custom AI processes are needed, the monthly costs add up for each new action that the user makes.

One of the main applications of these cloud-based AI's in consumer homes is integration with smart home devices such as light fixtures and smart tvs through a process called If This Then That (IFTTT). When a company wants to interface between their product and a smart home, they have to build the circuits and software to run them into the product itself. Usually, the company will also have the product connect to their servers to makes sure the software is up to date and has all the latest security patches to prevent backdoor access into the users home network. However, this not always free. Many companies charge a monthly fee to add IFTTT integration for their new device into their existing smart home environment and if a user is connecting most of the devices in their home that can be smart-enabled, then the many monthly fees can grow into a very large sum being paid each month.

benefits and drawbacks of hardware based and cloud based artificial intelligence

The main advantage of hardware over cloud based artificial intelligence is the increased power of the local hardware compared to the outsourced processing of a cloud based AI running on a remote server. This additional power lets the AI complete more complex tasks and work faster at tasks that both methods offer. On the other hand, the biggest advantage of a cloud based AI is the professional development of the commands that the user can give it. For example, the incredibly advanced feature that is Google Duplex has been developed by a whole team of software engineers from Google so the service is almost guaranteed to be of a high quality and be more polished than an open source project made by the community.

Another advantage of the hardware based AI is the use of it to supplement the existing hardware processing. In both the Huawei Mate 10 Pro and in Nvidia's RTX

2000 series graphics cards, Artificial Intelligence is used to increase the performance in the respective devices in a manner that will be noticeable to the end user, whether that is a more responsive mobile experience by reducing the performance drain of background tasks from the Huawei, or higher frames per second or better quality graphics in video games from offloading parts of the postprocessing to an AI specific part of the chip powered by Nvidia's graphics card.

But the hardware based AI lacks a level of portability and reproducibility that comes with using the same backend software on all of the same brand of cloud based AI. This means that, typically, a user can jump from device to device without having to relearn or reconfigure all of the features they had access to on the previous one which makes for a smoother user experience than writing custom AI's yourself.

An advantage of the cloud based AI is the low cost to entry for new users to get into using it. Because of this low cost, the user base is much larger which creates a bigger incentive for the companies who produce, update and maintain the cloud based AI's to keep doing what they are doing. As well as continuing, companies are encouraged to innovate in the fields surrounding AI which will help advance technology in the world as there is little reason for consumer targeted companies carry on innovating if there is no market willing to buy the final product from them.

This also works in reverse with the hardware side of AI. Currently, the chips required to power the AI's locally add a large amount onto the price of product they are applying the AI to. This means that fewer consumers are going to buy the product because of the higher price which lowers the incentive for the companies to continue development. This has a knock on effect for some companies such as Nvidia, whose products need regular software updates, in the form of device drivers, to add the performance improvements to new games which want to use the added features the AI brings to the table but if there aren't very many people using this technology, Nvidia are less likely to keep rolling out the updates to the AI for the various games that support its use.

conclusion

The main point in applying artificial intelligence to consumer devices is to improve the user experience. Be that through utilising the AI to improve performance in the short term or to help reduce stress to sensitive parts of the device like a battery that can otherwise shorten the lifespan of the device being in a usable state. The a large part of the consumer market are looking for a device that can perform better or longer rather than one where it's design is focused on looks. But if AI can reduce the performance hits and stop the shortening of the lifespan of a device caused by trying to improve the aesthetics then one device with performance, longevity and looks can be sold to all of the different parts of the market.

However, all of these benefits of AI typically come at a cost. And that cost is the price of the device. Devices with artificial intelligence often have a heavily marked up price due to the expense of developing the software, manufacturing the hardware and increasing the profit margins from the buzzword. But sometimes, when you find one that doesn't have an unreasonably inflated price for just reasons such as profit, then the increased cost doesn't seem to be as large of a pill to swallow for access to the newest technologies that will actually benefit you as a consumer.

But in reality, whether or not the application of Artificial Intelligence in consumer devices is good decision can only be decided by one thing. The consumers as a whole. If it is a good decision, then we'll see a rise in devices that utilise this technology over the next few years as consumer continue to buy the products featuring it. Otherwise, as the popularity of the AI dwindles, fewer and fewer devices will use it and it will fade into the obscurity of business applications until such time that the technology is mature enough to be affordable enough and powerful enough that it begins to make sense to apply it to the mainstream consumer devices.