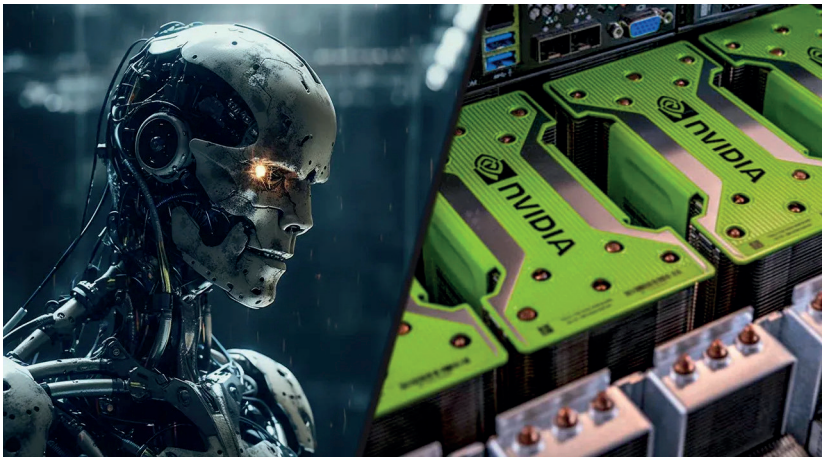
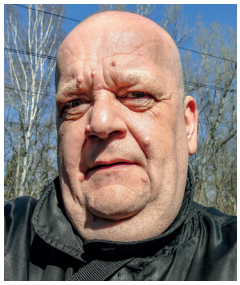
Sprzętowy wyścig do SI



Źródło: https://www.labellerr.com/blog/how-cuda-is-helping-tech-giants-to-train-foundation-models-10x-faster/

Minęły zaledwie dwa lata i kilka miesięcy od pierwszych publicznych prób ChatGPT, a rozmowy ze sztuczną inteligencją zdążyły trafić „pod strzechy" i dziś niemal do dobrego tonu należy zasięganie pomocy u cyfrowego geniusza. Do korzystania z ChatGPT i innych powstałych w międzyczasie źródeł przyznają się politycy i dziennikarze, nauczyciele narzekają na korzystanie ze wspomagania sztuczną inteligencją przez uczniów, a graficy i scenarzyści coraz częściej tracą intratne kontrakty. Prawdziwy wyścig do SI jednak dopiero się zaczyna, także w obszarze sprzętu.



Jacek Grabowski

z wykształcenia specjalista gazownictwa i górnictwa naftowego, przygodę z informatyką rozpoczął w końcu lat 80. XX wieku od współpracy z wydawnictwem „Lupus", gdzie publikował teksty głównie w dwutygodniku „PCkurier" i miesięczniku „Enter". Współtwórca pierwszego w Polsce informatycznego czasopisma B2B „MRK" (1997). Były redaktor naczelny miesięcznika „Reset", współpracownik wielu innych tytułów (magazyn „WWW", „IT Reseller", „Komputer Świat"). Obecnie freelancer, współpracuje m.in. z warszawską komunikacją miejską.

Efekt, jaki wywołało otwarcie przed ludźmi czatów czy generatorów obrazów wspomaganych przez cyfrowe systemy eksperckie, przyniósł ze sobą olbrzymi impuls dla biznesu.

Za i przeciw

Sztuczna inteligencja jest obecnie w szczycie zainteresowania opinii publicznej, mając zarówno zagorzałych przeciwników, jak i zwolenników. Jednym z bardziej znanych „wrogów" SI w Polsce jest socjolog prof. Andrzej Zybertowicz. Choć, moim zdaniem, jego sugestie o możliwym wykształceniu samoświadomości w gęstwinie sieci neuronowej idą zbyt daleko, to jednak warto zwrócić uwagę, że w popkulturze pojawił się i funkcjonuje taki mit, bo rzekomo „informatycy sami nie wiedzą, jak to w ogóle działa". Być może nawet ten mit jest po cichu wspierany marketingowo i ma podgrzewać zainteresowanie zagadnieniem „cudu", jakim jest rozmowa z napakowanym danymi botem napędzanym setkami kart graficznych.

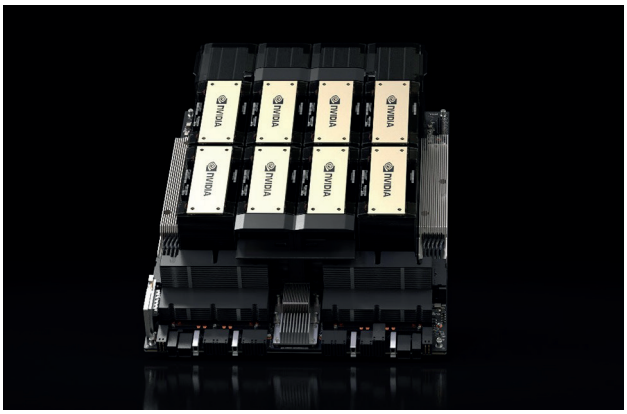
Jednak nie wszystkie zastrzeżenia Zybertowicza łatwo zignorować i uznać za narzekania człowieka starszego pokolenia, który nie rozumie nowoczesności i informatyki. Wskazuje on bowiem na różne faktycznie negatywne skutki społecznego „kultu" ChatGPT i innych tego typu rozwiązań, jak np. możliwość „podpowiadania" ludziom pewnych zachowań i poglądów, udzielanie rad na temat odżywiania czy stylu życia. Zybertowicz wskazuje też na SI jako „mnożnik mocy" człowieka, a bez wątpienia taki mnożnik może służyć zarówno celom dobrym, jak i złym, np. ataki hakerskie czy rozpowszechnianie dezinformacji. Ludzie zbyt łatwo wierzą w osądy rozreklamowanych systemów eksperckich, które potrafią także zgrabnie wprowadzać w błąd, co – według Zybertowicza – może spowodować wiele perturbacjii i prowadzić do ogłupiania ludzi zamiast dostarczania im prawdziwej wiedzy.

Pierwsze szkodliwe skutki rozpowszechnienia takiej inżynierii wiedzy są już zauważalne na rynku pracy. Na przykład ministra Agnieszka Dziemianowicz-Bąk ogłosiła, że rząd ma opracować listę zawodów, których nie wolno będzie zastępować robotami. Przyczyniła się do tego zapewne decyzja kierownictwa stacji RadioOFF w Krakowie o eksperymentalnym zastąpieniu dziennikarzy sztuczną inteligencją. Po miażdżącej fali krytyki, która przetoczyła się we wszystkich mediach, z eksperymentu wycofano się po zaledwie tygodniu. To jest dopiero początek takich przedsięwzięć, więc na razie ludzie obawiający się utraty pracy są raczej solidarni w sprzeciwie, a rozwiązania robotyczne wciąż na początku swojej drogi.

Niezależnie jednak od tego, że systemy eksperckie mogą istotnie stanowić niebezpieczeństwo dla niektórych zawodów, entuzjaści SI, porównując pomysł rządu do angielskiego społecznego ruchu tkaczy niszczących na początku XIX w. maszyny stanowiące dla nich zbyt silną konkurencję, nazywają działania przeciw rozpowszechnianiu sztucznej inteligencji „nowym luddyzmem". Trudno odmówić racji ich argumentom, że rozwoju systemów SI różnego rodzaju nie da się w żaden sposób powstrzymać, a nienaturalne hamowanie rynku decyzjami rządowymi przyczyni się raczej do wzrostu zacofania technologicznego, niż przyniesie dobre skutki społeczne. Istotnie – rozwój własnych technologii inżynierii wiedzy jest obecnie kluczowym tematem, każdy chce coś ukroić dla siebie z tego tortu. A świat, nie zważając na zastrzeżenia i problemy społeczne, galopuje. Warto wspomnieć, że od niedawna pojawiła się możliwość prowadzenia płynnej rozmowy z ChatGPT, nic nie trzeba wpisywać, a robot odpowiada „jak człowiek", poprawnie formułując zdania i nawiązując konwersację. Prof. Zybertowicz ostrzega, że tylko czekać, jak ludzie zaczną prowadzić intymne rozmowy z robotami, które mają szansę zastąpić nocne audycje radiowe i będą udzielać porad życiowych.

CUDA się dzieją

Do tej pory liderem rynku sprzętowych rozwiązań SI była NVIDIA ze swoimi GPU – procesorami graficznymi. Źródłem sukcesu GPU tej firmy stała się opracowana w 2007 r. architektura CUDA (Compute Unified Device Architecture), stanowiąca połączenie specyfikacji masowo równoległego procesora wielordzeniowego z API niezależnym od sprzętu, przez co napisane raz programy można uruchamiać na coraz nowszych kartach w tym standardzie, nie ma problemu ze zmianą programów obsługi itd. Pierwszym masowym zastosowaniem GPU do obliczeń innych niż grafika były kryptowaluty i algorytm blockchain, potem karty graficzne w pecetach zaczęły być coraz częściej wykorzystywane przez konsumenckie edytory wideo do przyspieszania renderowania filmów (dziś to właściwie standard), aby wreszcie opanować rynek sztucznej inteligencji.



Jednostka graficzna NVIDIA H200 Tensor Core GPU – pamięć 141 GB, przepustowość pamięci 4,8 TB/s, teoretyczna wydajność FP84 petaFLOP, TPD do 1000 W

Źródło: https://www.nvidia.com/en-us/data-center/h200/

Jednak wiele wskazuje jednak na to, że bieżący rok może być przełomowy, jeśli chodzi o wspomaganie przetwarzania sieci neuronowych przez procesory komputerów osobistych. Warto przypomnieć, że gdy w 1993 r. pojawiła się metoda kompresji dźwięku, znana potocznie jako MP3, pecety nie były zdolne do odtwarzania tak zapisanych plików. W 2001 r. masowo już ściągano je z internetu i słuchano na domowych komputerach, a na rynku pojawił się pierwszy zminiaturyzowany odtwarzacz przenośny – iPod Apple'a. Rozwój wydarzeń nasuwa skojarzenie, że ze sztuczną inteligencją będzie podobnie, tylko znacznie szybciej.

> *Do tej pory myśląc o SI wyobrażaliśmy sobie raczej jakieś farmy serwerów z wieloma kartami graficznymi pożerającymi mnóstwo energii, tymczasem właśnie w tym roku po raz pierwszy szczególnie zaznaczył się trend poszukiwania tańszych i łatwiej dostępnych alternatyw sprzętowych do wspomagania obliczeń sieci neuronowych.*

NPU – nowy procesor, nowe obliczenia

Już od kilku lat można odnotować przy opisach procesorów pojawienie się nowego skrótu – NPU. Kryje się pod nim Neural Processing Unit (jednostka przetwarzania neuronowego), której sama nazwa wskazuje na optymalizację architektury pod kątem przetwarzania sieci neuronowych i aplikacji głębokiego uczenia się oraz uczenia maszynowego. NPU nie zostały wymyślone w celu bezpośredniego konkurowania z CPU czy GPU, są raczej rodzajem koprocesora zintegrowanego z nimi w jedną platformę. Cechą wyróżniającą NPU jest przeznaczenie do przetwarzania równoległego czy też systolicznego (łączącego przetwarzanie równoległe i potokowe). Podczas gdy tradycyjne procesory wymagają wykonania wielu instrukcji do zakończenia przetwarzania neuronów, jednostka NPU jest w stanie wykonać zbliżoną operację za pomocą tylko jednej instrukcji. Oczywiście procesory graficzne (GPU) są podobnie wydajne w przetwarzaniu równoległym, jednak dzięki innej architekturze NPU może przewyższać równoważny procesor graficzny przy zmniejszonym zużyciu energii i mniejszej powierzchni fizycznej, ponieważ – w przeciwieństwie do bardziej uniwersalnego GPU – został on zaprojektowany wyłącznie dla zastosowań w obliczeniach charakterystycznych dla sieci neuronowych.

Pojawienie się zapotrzebowania na dodanie NPU do konsumenckich platform pecetowych pociągnęło za sobą cały rynek procesorów, bowiem „wszczepienie" jednostki neuronowej do jednostki centralnej, zintegrowanej z procesorem graficznym wymusiło dalsze zmiany technologiczne. Spopularyzowanie systemów eksperckich stało się impulsem do nowej wojny między gigantami technologicznymi i zarazem początkiem kolejnego mnożenia mocy obliczeniowej i walki na FLOPS-y. Teraz walka przenosi się raczej na teren NPU, którego moc w uproszczony sposób mierzona jest w jednostkach TOPS, czyli Tera Operations Per Second (biliony operacji na sekundę).

Apple przewodzi

Jednym z pierwszych procesorów wyposażonych w NPU był Apple A11, którego premiera miała miejsce we wrześniu 2017 r. Zintegrowany układ oferował moc mniej więcej na poziomie 0,6 TOPS. Firma Apple konsekwentnie udoskonalała produkt, wobec czego wbudowany NPU w 2020 r. osiągnął już 11 TOPS (w procesorze A14), rok później w A15 ta wydajność jeszcze wzrosła, osiągając prawie 16 TOPS. We wrześniu 2023 r. pojawił się A17 Pro, którego NPU osiągnął już 35 TOPS. Niecały rok później, w maju 2024 r., NPU w procesorze Apple M4 doszedł do 38 TOPS. Także inni producenci procesorów, głównie smartfonowych, od kilku lat oferują NPU wspierające funkcje AI w ich telefonach, osiągające obecnie podobne wyniki jak jednostki Apple'a.

Na tle tych osiągnięć przez długi czas blado wypadali producenci procesorów do komputerów PC. Wydaje się, że początkowo lekceważyli konsumencki rynek rozwiązań bazujących na inżynierii wiedzy, koncentrując się głównie na oferowaniu coraz szybszych CPU dla graczy czy „górników" kryptowalut. Najwyraźniej zadziałał na nich dopiero efekt premiery ChatGPT, bo pierwsze procesory Intela i AMD z zintegrowanym NPU pojawiły się na rynku w 2023 r. i miały jednostki neuronowe o mocy – odpowiednio – 11 i 16 TOPS. W pecetach NPU pojawiły się sześć lat później niż u Apple'a i innych producentów, w dodatku w momencie premiery były zauważalnie słabsze od już istniejących rozwiązań konkurencji.

Same liczby nie obrazują dokładnie, co potrafi NPU. Dla porównania przyjrzyjmy się najszybszej obecnie karcie graficznej dla graczy Nvidia RTX 4090, wykorzystującej procesor (GPU) o oznaczeniu AD102 zawierający 11 bloków GPC, w których mieszczą się: 16 384 procesory CUDA, 512 jednostek teksturujących, 128 jednostek renderujących, 128 rdzeni RT oraz 512 tensorów. Dzięki zastosowaniu nowej litografii 5 nm tajwańskiej firmy TSMC, która zastąpiła stosowane wcześniej 8 nm Samsunga, w nowym GPU niemal trzykrotnie zwiększono liczbę tranzystorów przy jednoczesnym zmniejszeniu rozmiarów rdzenia. Taka karta, kosztująca ok. 11 tys. złotych i zużywająca sporo energii (TDP 450 W)[1], osiąga ponad 1000 TOPS. Na pierwszy rzut oka NPU nie wydają się specjalnie konkurencyjne wobec rozwiązań opartych na GPU, choć wyróżniają się zmniejszonym zapotrzebowaniem na energię (TDP całej platformy, czyli CPU/GPU/NPU, waha się w granicach 125–180W), ale to wciąż dopiero początek rozwijania tej koncepcji. Poza tym obecność NPU nie znaczy, że komputer osobisty stanie się od razu silnikiem do napędzania własnego czata i zastąpi internetowe usługi.

Copilot+PC i co z tego wynikło?

W maju tego roku Microsoft wspólnie z producentem procesorów Qualcomm opublikowali specyfikację przeno-

1 TPD (Thermal Design Power) to maksymalną ilość ciepła, jaką układ może wygenerować w warunkach typowego obciążenia. TDP jest proporcjonalne do zużycia energii przez układ.

śnych komputerów osobistych ze wspomaganiem przetwarzania sieci neuronowych w systemie Windows 11 nazwaną Copilot+PC. Copilot (w lotnictwie oznacza „drugiego pilota") to nazwa aplikacji wbudowanej w system Windows, będącej interfejsem do czatu ze sztuczną inteligencją. Komputer w specyfikacji Copilot+PC to jednak coś więcej niż samo czatowanie, bowiem nowa wersja Windows 11 oferuje kilka funkcji napędzanych sieciami neuronowymi, np. Automatyczną Super Rozdzielczość, czyli poprawianie jakości obrazu w czasie rzeczywistym; wbudowany w program malarski Paint efekt Cocreator, czyli generator grafiki SI; wspomagane sztuczną inteligencją efekty w edytorze fotografii Windows Studio i jeszcze kilka drobiazgów. Ponieważ wszystko to wymaga odpowiedniego sprzętu, Microsoft z Qualcommem ustalili pewne wymagania, które ma spełnić notebook w standardzie Copilot+PC.

Najważniejszym zaleceniem nowej specyfikacji jest wyposażenie notebooka w procesor ze zintegrowanym układem NPU. Bez niego wspomniane efekty i dodatki albo w ogóle nie będą działać, albo będą z nimi kłopoty. Moc wymaganego NPU została określona i ma przekraczać 40 TOPS. Poza tym komputer musi mieć minimum 16 GB pamięci operacyjnej i dysk SSD lub pamięć flash UFS o pojemności minimum 256 GB. Ostatnim wymogiem uzyskania certyfikatu Copilot+PC dla komputera jest umieszczenie na klawiaturze dodatkowego przycisku wywołującego bezpośrednio aplikację Copilot.

W momencie ogłaszania nowej specyfikacji tylko jeden producent procesorów dysponował produktem spełniającym wymaganie mocy NPU. Nie będzie tajemnicą, że to właśnie Qualcomm, czyli wspólnik Microsoftu – jego procesory Snapdragon X oraz Snapdragon X Elite dysponowały bowiem rozwiązaniem Hexagon NPU o mocy 45 TOPS, szybszym nawet od Apple'a. Wprawdzie Snapdragon jako procesor w architekturze ARM64 kojarzy się bardziej ze smartfonami, ale to nie jest problem, gdyż może również bez problemów współpracować z Windows 11, a jest to naprawdę silny i energooszczędny układ (TDP 25 W). Pomijając marketingowy aspekt Copilot+PC, wprowadzenie komputerów spełniających wymagania tej specyfikacji na rynek PC to znaczący krok naprzód we wszczepianiu NPU do platformy pecetowo-windowsowej i kolejny impuls dla rynku procesorów, który żywi się takimi wyzwaniami. Postawienie progu mocy NPU, która promowała procesory Qualcomm, było przecież „kopem" dla innych producentów, aby dogonić i przegonić Snapdragona od pecetów.

Lunar Lake i AI300

Niedługo po ogłoszeniu specyfikacji Copilot+PC Intel zaprezentował drugą generację procesorów mobilnych z wbudowanym NPU, nazwaną Lunar Lake. Na rynku funkcjonują one jako rodzina Intel Core Ultra 200V. Do ich produkcji TSMC użyło nowego procesu N3B (3 nm), ale nie jest to jak dawniej jeden wafel krzemowy, a raczej wafelek przekładany, gdzie mamy jakby kilka osobnych płytek upakowanych metodą Foveros 3D. Poszczególne płytki mają różne zadania – układ Compute Tile zawiera rdzenie CPU o największej wydajności, Graphics Tile – procesor graficzny, IO Tile – układy we/wy, a SoC Tile zawiera NPU oraz rdzenie o niższej wydajności zastępujące rdzenie na płytce Compute Tile przy mniej wymagających zadaniach, co oszczędza energię. Razem z wprowadzeniem Lunar Lake Intel zaproponował także podział wyników wydajności mierzonych w TOPS na PTOPS (wynik całej platformy) i pTOPS (wynik samego NPU). W przypadku Lunar Lake prodcent podaje wartość 120 PTOPS i 48 pTOPS. Mimo zabiegów marketingowych wszystkie testy pokazują jednak, że Lunar Lake przegrywa z nową rodziną mobilnych procesorów AMD, znanych jako AI300.

Procesory AI300 pojawiły się wkrótce po ogłoszeniu Copilot+PC. Na przykład Ryzen AI 9 HX 370 pochodzący z tej rodziny ma 12 rdzeni (w tym cztery o wysokiej wydajności i osiem słabszych) oraz zintegrowany układ graficzny AMD Radeon 890M z 16 blokami CU RDNA 3.5. Słabszy model Ryzen AI 9 365 ma tylko sześć rdzeni niższej wydajności. Producent nie zapomniał także o NPU, który w obu pro-cesorach ma osiągać 50 TOPS, czyli więcej niż Intel. Nie tylko na tym polu Intel okazuje się słabszy, w niektórych benchmarkach najsilniejsza platforma Ryzena przewyższa najnowsze Core Ultra prawie dwukrotnie, jednak dzieje się to kosztem nieco większego zużycia energii (Intel podaje typowe TDP 23 W, szczytowe 37 W, natomiast AMD odpo-wiednio 28 W i 54 W.

TPU – jeszcze inny procesor SI

W 2015 r. firma Google dla własnych potrzeb opracowała procesor wspomagający obliczenia sieci neuronowych, nazwany Tensor Processing Unit (TPU). Później TPU zaczęły być udostępniane szerszemu gronu użytkowników, m.in. w smartfonach Google Pixel. TPU różnią się zasadniczo od NPU, które są oparte na tradycyjnej architekturze von Neumanna oddzielającej pamięć operacyjną od procesora. Zamiast tego, w TPU pamięć jest zintegrowana z rdzeniami obliczeniowymi. TPU pracuje w trybie przetwarzania systolicznego (dosłownie „skurczowego"). Procesor rytmicznie wykonuje operacje obliczeniowe (w sposób przypominający skurcze serca, skąd nazwa) na wektorach danych uformowanych w potoki. Jednostki TPU mają niższą wydajność szczytową niż NPU, co oznacza, że mogą przetwarzać raczej specyficzne, zoptymalizowane sieci neuronowe. Z drugiej strony charakteryzują się niższą latencją i zużyciem energii niż TPU. TPU produkuje tylko Google, więc na rynku nie są tak często spotykane, jak NPU.