

Uniwersytet Jagielloński trenuje sieci neuronowe z wykorzystaniem technologii NVIDIA GPU

Wydział Matematyki Uniwersytetu Jagiellońskiego stworzył załączek infrastruktury opartej na technologii NVIDIA GPU. Na platformie trenowane są sieci neuronowe, których wykorzystanie istotnie wpłynie na szybkość i jakość działania m.in. takich branż jak chemioinformatyka, kryminalistyka czy rozwój sektora pojazdów autonomicznych.

POTRZEBA KLIENTA

Grupa badawcza naukowców Uniwersytetu Jagiellońskiego (GMUM – Group of Machine Learning Research) specjalizująca się w uczeniu maszynowym, w szczególności głębokim uczeniem się - zarówno w kontekście podstawowym, jak i stosowanym, pod kierunkiem prof. Jacka Tabora, potrzebowała stworzyć załączek infrastruktury obliczeniowej opartej o GPU. Wykorzystanie GPU mogło zagwarantować grupie platformę do realizacji badań z zakresu: tworzenia modeli generatywnych, teoretycznego rozumienia i optymalizacji procesów głębokiego uczenia się, przetwarzania języków naturalnych, projektowania struktur leków (chemioinformatyka) czy analizy komputerowej obrazów medycznych.

Patrząc na ilość zadań obliczeniowych, z jakimi mamy do czynienia, zdefiniowaliśmy, że klaster powinien mieć 16 GPU. Dodatkowo wszystkie karty graficzne powinny być dostępne do obliczeń równoległych. W przypadku zadań uruchamianych jednocześnie w obrębie jednego klastra, wysokowydajne interfejsy sieciowe powinny pozwalać na pełną użycie GPU. Dlatego zdecydowaliśmy się na rozwiązania NVIDIA – mówi prof. Jacek Tabor z Uniwersytetu Jagiellońskiego.

ROZWIĄZANIE

Ze względu na specyfikę prowadzonych badań serwer został wyposażony w cache dużej pojemności pozwalając na lokalne przetwarzanie danych podczas procesu uczenia sztucznych sieci neuronowych. Sam klaster zaprojektowano tak, aby możliwa była łatwa rozbudowa o kolejne węzły. Stworzony przez NVIDIA ekosystem pozwala na szybkie budowanie rozwiązań opartych o wszystkie popularne frameworki. Mnogość bibliotek matematycznych przyspiesza obliczenia przenosząc ich ciężar z klasycznych procesorów na jednostki GPU.

„Rozwiązanie zostało zbudowane w metodyce switchless, co pozwoliło na spełnienie kluczowych wymagań. Całość składa się z dwóch serwerów NVIDIA DGX-1, każdy z nich dysponuje ośmioma kartami NVIDIA TESLA V100, które pozwalają na osiągnięcie wydajności obliczeniowej na poziomie 1 petaFLOPS. Serwery zostały połączone ze sobą bezpośrednio za pomocą 4 połączeń 100 Gb InfiniBand, co bez przeszkód pozwala w pełni użycie klaster nawet w przypadku najbardziej intensywnych zadań. Każdy z węzłów został również wyposażony w dwa interfejsy 10GbETH pozwalające na łatwe podłączenie przestrzeni dyskowych dostępnych w środowisku Uniwersytetu Jagiellońskiego” – mówi Jakub Zboina, AI Platform Architect, Comtegra S.A.

Zespół naukowców mógł przystąpić do pracy od razu po instalacji klastra, co było dla nas niezwykle istotne.

- 2 serwery NVIDIA DGX-1, posiadające każdy po 8 kart NVIDIA TESLA V100
- Wydajność obliczeniowa na poziomie 1 petaFLOPS
- Serwery połączone 4 połączeniami 100 Gb InfiniBand
- Każdy węzeł wyposażony w dwa interfejsy 10GbETH

EFEKTY

„Dostarczona przez Comtegra infrastruktura to w pełni rozbudowywalna, ale przede wszystkim funkcjonalna, szybka i optymalna platforma dla naszego zespołu naukowego. Zadania, z jakimi mierzy się zespół, ich mnogość, ilość obliczeń wykonywanych jednocześnie – to jest ogromne wyzwanie. Ale wiemy, że dzięki modernizacji platformy i wykorzystaniu technologii NVIDIA GPU jesteśmy w stanie naprawdę realnie wspierać rozwój przemysłu farmaceutycznego, organów ścigania i wielu innych branż dostarczając im gotowe rozwiązania oparte na sztucznej inteligencji” – mówi prof. Jacek Tabor z Uniwersytetu Jagiellońskiego.

Dzięki implementacji klastra wyposażonego w akceleratory naukowcy z Uniwersytetu Jagiellońskiego byli w stanie przyspieszyć badania o ok. 47 razy względem używanej poprzednio architektury opartej o CPU. Co znaczy, że trening sieci neuronowej trwający jeden dzień został skrócony do 30 minut używając tylko jednego z 16-stu akceleratorów. Taki wzrost wydajności pozwala na szybsze testowanie różnych modeli przez co zauważalnie skraca się czas potrzebny na znalezienie optymalnego rozwiązania.

- Badania przyspieszono 47 razy w stosunku do czasu, jaki potrzebowała architektura oparta o CPU
- 1 dzień obliczeń na CPU = 30 minut z wykorzystaniem GPU
- Oszczędność energii elektrycznej
- Oszczędność miejsca w datacenter

Czas oczekiwania na wyniki realizowanego zadania to nie tylko koszty związane z utrzymaniem pracownika czy wydłużony proces dostarczenia gotowego produktu. Znacznie szybsze wykonywanie się zadań wpływa również na niższe zużycie energii elektrycznej oraz brak konieczności budowania rozwiązań sprzętowych nierzadko zajmujących kilkanaście szaf rackowych. To z kolei przekłada się bezpośrednio na niższe koszty utrzymania Data Center.

Przed kupnem omawianego rozwiązania Uniwersytet stanął również przed innym problemem. Wyposażeni jedynie w konsumenckie karty graficzne, znajdujące się w komputerach osobistych dostępnych w laboratoriach, naukowcy nie byli w stanie przenieść do pamięci kart graficznych nawet jednego pełnego zdjęcia badanych obiektów. Zdjęcia o bardzo dużej rozdzielczości nie mieściły się bowiem w pamięci podręcznej kart przeznaczonych do renderowania grafiki co nie pozwalało na przetwarzanie ich jako całości i stanowiło dużą przeszkodę w dalszym kontynuowaniu badań. Dzięki podwojeniu dostępnej pamięci oraz zwiększeniu jej przepustowości poprzez zastosowanie akceleratorów NVIDIA V100 naukowcy byli w stanie w prosty sposób skrócić czas przygotowywania surowych danych. Co pozwoliło na jeszcze szybsze prowadzenie badań – podsumowuje Jakub Zboina, AI Platform Architect, Comtegra S.A.

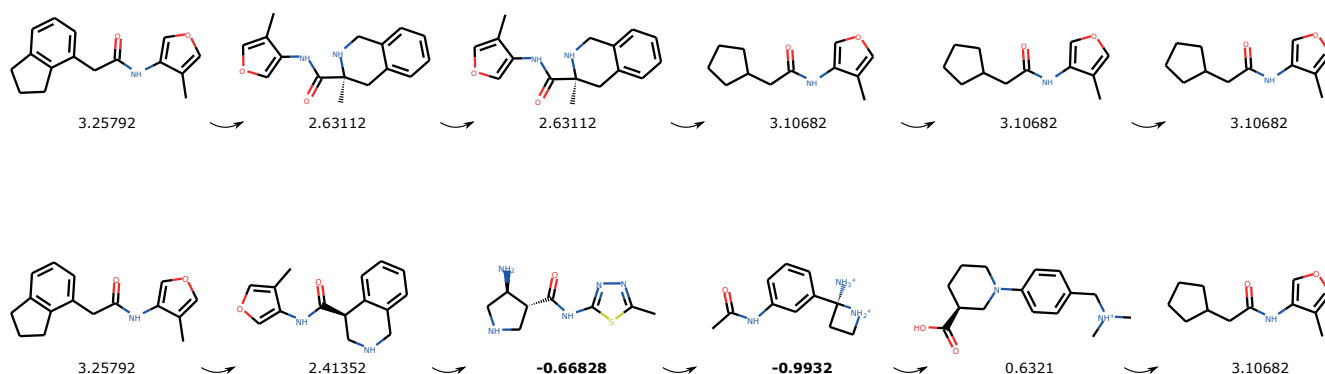


Interpolacja oparta na sieciach neuronowych w ukrytych przestrzeniach

Łukasz Struski, Michał Sadowski, Tomasz Danel, Igor Podolak, prof. Jacek Tabor, Wydział Matematyki i Informatyki

Zastosowanie praktyczne: Znajdowanie związków chemicznych o określonych właściwościach (na przykład w celu stworzenia nowych leków). Inteligentna edycja zdjęć (np. w celu wsparcia pracy organów ścigania w procesie tworzenia wersji potencjalnych portretów długo poszukiwanych przestępców).

Modele generatywne zazwyczaj generują nowe obrazy z losowych punktów z tzw. Przestrzeni ukrytej. Na liniowej ścieżce między dwoma punktami w ukrytej przestrzeni można utworzyć serię punktów. Następnie punkty te można wykorzystać do wygenerowania serii obrazów, które pokazują przejście między dwoma wygenerowanymi obrazami. Przy dwóch obrazach taka metoda pozwala na generowanie nowych obrazów, które są do nich w pewnym stopniu podobne. Zespół opracował metodę wymuszającą, tak, że generowane obrazy spełniają dodatkowo pewien zestaw predefiniowanych wymagań. Przykładowo, mając dwa obrazy twarzy, możemy wygenerować nowe twarze, które są do nich podobne i dodatkowo zawierają np. okulary, inny kolor włosów, brodę. Model działa nie tylko na zdjęciach. Mając na przykład chemiczny model generatywny (generujący wykresy cząsteczek chemicznych), możemy zapewnić, że wszystkie punkty na ścieżce odpowiadają związkom lekopodobnym. Taka metoda mogłaby posłużyć do odkrycia nowych leków o określonych cechach.



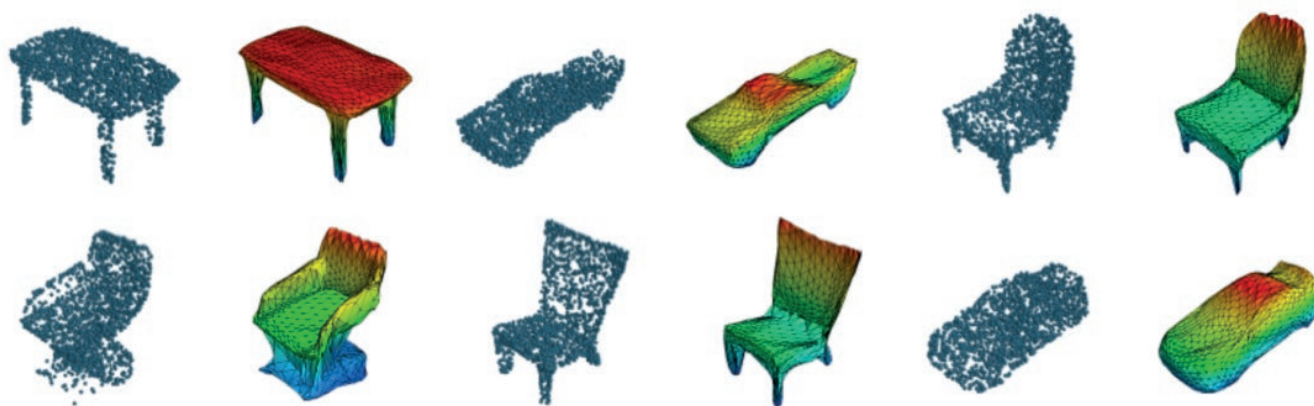
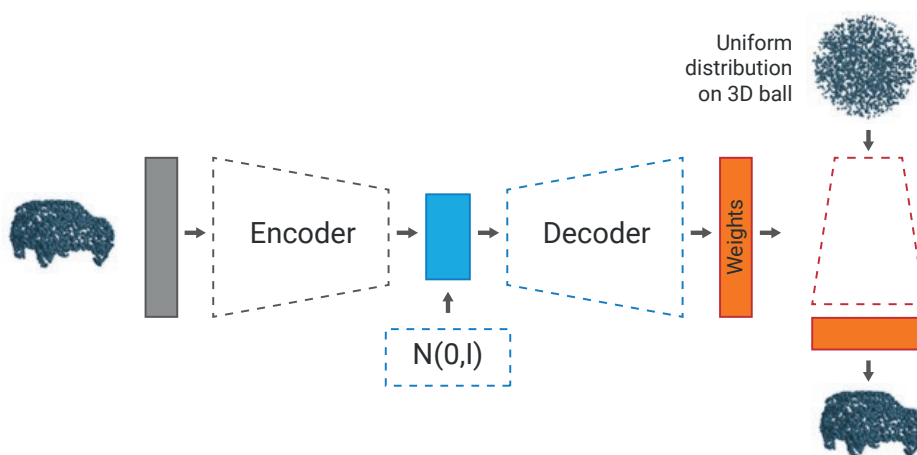
Podejście hipersieciowe do generowania chmur punktów

Przemysław Spurek, Sebastian Winczowski, Jacek Tabor, Maciej Zamorski, Maciej Zięba, Tomasz Trzciński

Zastosowanie praktyczne: Zwiększenie efektywności rozpoznawania obiektów 3d, z wykorzystaniem radarów (używanych np. w samochodach autonomicznych).

Hypernetwork to podejście polegające na wykorzystaniu jednej sieci neuronowej do generowania wag dla innej (zwykle większej) sieci neuronowej. W tym projekcie zespół wykorzystał tę metodę do generowania obiektów 3D.

Sieci neuronowe można łatwo wyszkolić do generowania chmur 3D o stałej liczbie punktów. Niestety nie można ich użyć do wygenerowania gęstszych chmur punktów, ani oszacowania ich powierzchni. Funkcje hipersieci mogą rozwiązać te problemy. Metoda pozwala hipersieci generować wagi sieci docelowej, które można zinterpretować jako parametryzację powierzchni kształtu 3D. Dokładniej, zadaniem sieci docelowej jest odwzorowanie jednolitej kuli punktów na zadany kształt. W konsekwencji sfera jednostkowa zostaje przekształcona w granicę obiektu.



Strukturalne sztuczne sieci neuronowe

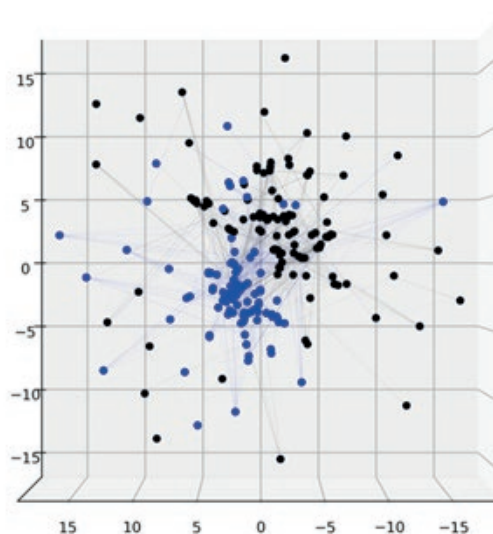
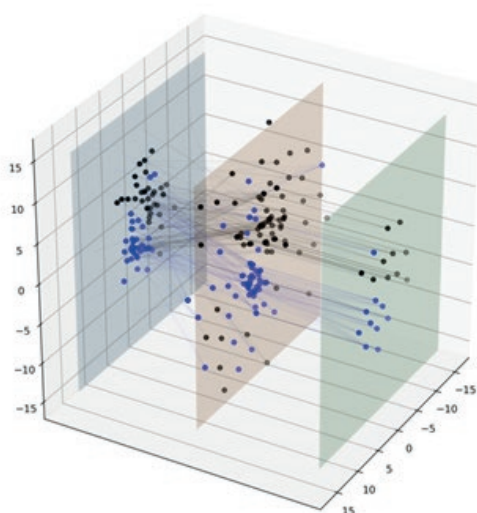
Maciej Wołczyk, Jacek Tabor, Marek Śmieja, Szymon Maszke

Zastosowanie praktyczne: Tworzenie małych sieci neuronowych (na przykład na urządzenia mobilne) zdolnych do rozwiązywania różnego typu problemów.

W tym projekcie wprowadzamy inspirowane biologicznie sztuczne sieci neuronowe składające się z neuronów, które dodatkowo charakteryzują się położeniami przestrzennymi. Aby zasymulować właściwości systemów biologicznych, dodajemy koszty obciążające długie połączenia i bliskość neuronów w dwuwymiarowej przestrzeni.

Eksperymenty pokazują, że w przypadku, gdy sieć wykonuje dwa różne zadania, neurony w naturalny sposób dzielą się na klastry, z których każdy jest odpowiedzialny za przetwarzanie innego zadania. To zachowanie nie tylko odpowiada systemom biologicznym, ale także pozwala na dalszy wgląd w interpretowalność lub ciągłe uczenie się.

Badamy regularizujące właściwości naszego modelu. W szczególności sieć zwykle wypycha wszystkie neurony, które nie są ważne dla rozwiązania zadania, ponieważ trzymanie ich blisko utworzonych klastrów dodaje dodatkowe wartości strat. W praktyce oznacza to, że sieci o niskim potencjale można odciąć od sieci, zapewniając w efekcie bardzo naturalny sposób przycinania sieci.



Visualization of neurons' positions in the neural network.



O Uniwersytecie Jagiellońskim

Uniwersytet Krakowski jest najstarszą uczelnią w Polsce i jedną z najstarszych w tym regionie Europy. Został fundowany 12 maja 1364 r. przez króla Kazimierza Wielkiego. Studium generale – tak wówczas nazywano Uniwersytet – składało się z trzech wydziałów: sztuk wyzwolonych, medycyny i prawa.

Obecnie UJ to ponad 540 profesorów, 730 doktorów habilitowanych, 2600 osób pozostałej kadry nauczycielskiej, około 50 tysięcy studentów i doktorantów oraz ponad 3,5 tysiąca pracowników administracji. W chwili obecnej Uniwersytet Jagielloński posiada 16 wydziałów, w tym trzy medyczne. Rokrocznie kształci 50 tysięcy studentów, z czego 65% stanowią kobiety.

Uniwersytet stale zmienia swoje oblicze i modernizuje szeroko pojętą infrastrukturę. Inwestycje takie jak przeniesienie Wydziału Fizyki, Astronomii i Informatyki Stosowanej oraz Wydziału Chemii do ukończonej nowej siedziby, budowa Centrum Edukacji Przyrodniczej i Instytutu Geologii, to tylko niektóre z nich. Modernizacje i rozbudowy tworzą przyjazne miejsce pracy naukowców i uniwersyteckiej kadry nauczycielskiej oraz przyjazną strefę do nauki dla studentów.



O Comtegra

Comtegra jest integratorem systemów informatycznych specjalizującym się w obszarach: Data Protection & Business Continuity, Data Management, Cloud Computing, Networking, Cybersecurity oraz HPC i sztucznej inteligencji. Firma świadczy także usługi w modelu Cloud. Obecnie Comtegra to stabilna firma, z ponad 20-letnim doświadczeniem, tworzona przez zespół prawie 100 specjalistów, których kompetencje pozwalają realizować najbardziej zaawansowane projekty teleinformatyczne.

Miejsce Comtegra na rynku polskich podmiotów działających w branży ICT odzwierciedlają zarówno wysokie statusy partnerskie światowych dostawców technologii ICT, jak i pozycje zajmowane przez firmę w niezależnych rankingach branżowych. Obecnie Comtegra to wiodący partner sprzedażowy m.in.: Dell Technologies #1, VMware #1, Veritas #1, NetApp #1, Cisco #2 (obszar rozwiązań DataCenter) oraz Broadcom TOP3. Więcej informacji o firmie można znaleźć pod adresem: www.comtegra.pl

Chcesz dowiedzieć się więcej? Odwiedź nasze strony:

Rozwiązania oparte na sztucznej inteligencji dla biznesu i nauki



Rozwiązania dla centrów superkomputerowych



Comtegra S.A.

Ul. Puławska 474, 02-884 Warszawa

tel.: +48 22 311 18 00

info@comtegra.pl

www.comtegra.pl

Partner publikacji

