Odkrywanie większej ilości sztucznej inteligencji w Pythonie

Po przeczytaniu poprzednich trzech Części dowiedziałeś się całkiem sporo o korzystaniu z niektórych podstaw sztucznej inteligencji, w szczególności sieci neuronowych i uczenia maszynowego. Sztuczna inteligencja to jednak znacznie więcej niż tylko te dwie rzeczy. Moglibyśmy przyjrzeć się wyszukiwaniu zaawansowanemu (nie wyszukiwaniu w Google, ale raczej przyglądaniu się dużym przestrzeniom problemowym i próbowaniu znalezienia rozwiązań problemu przy użyciu sztucznej inteligencji). Moglibyśmy również przyjrzeć się całemu problemowi autonomii robotyki, ale ten temat jest bardzo skomplikowany. Zamiast tego porozmawiamy o innych sposobach tworzenia oprogramowania AI poza Raspberry Pi. Pamiętasz, jak siedem godzin zajęło nam przeprowadzenie pięciu epok uczenia w naszej dużej sieci neuronowej? Wygląda na to, że moglibyśmy użyć większego żelaza, aby wykonać więcej treningów w krótszym czasie.

Ograniczenia Raspberry Pi i AI

Raspberry Pi to niedrogie pełnowymiarowe urządzenie komputerowe. Raspberry Pi 3B, z którego korzystaliśmy w tej książce, ma następujące główne specyfikacje:

Procesor: Czterordzeniowy 64-bitowy procesor Broadcom @ 1,4 GHz

Karta graficzna: Broadcom Videocore-IV

Pamięć RAM: 1 GB pamięci SDRAM

Sieć: Gigabit Ethernet, 802.11b/g/n/ac Wi-Fi

Przechowywanie: karta SD

Jak to się układa? Za komputer za 35 dolarów, bardzo dobrze. Ale dla dedykowanego AIkomputera, nie tak bardzo. Problemy to niewystarczająca ilość pamięci RAM (1 GB to niewiele, zwłaszcza dla Raspberry Pi do sztucznej inteligencji) i niezbyt wyrafinowany procesor graficzny (jednostka przetwarzania grafiki). Istnieją dwie okoliczności łagodzące, które utrzymują Raspberry Pi w działaniu, jeśli chodzi o robienie i eksperymentowanie z AI. Po pierwsze, możesz kupić akcelerator AI, który można podłączyć do portów USB Raspberry Pi, a po drugie, możesz używać Raspberry Pi do sterowania procesorami i sprzętem AI w chmurze w celu wykonania wszystkich zadań wymagających dużej mocy obliczeniowej. Pamiętaj z naszego poprzedniego rozdziału, że większość czasu komputera w budowaniu dowolnego systemu uczenia maszynowego sztucznej inteligencji jest przeznaczona na szkolenie, a kiedy to szkolenie jest zakończone, nie wymaga dużo przetwarzania, aby faktycznie scharakteryzować nieznany lub nowy obraz . Oznacza to, że możesz trenować na jednej dużej maszynie, a następnie wdrożyć aplikację na prostszym komputerze, takim jak Raspberry Pi. To nie działa przez cały czas (zwłaszcza jeśli chcesz kontynuować naukę w trakcie działania programu), ale jeśli działa, pozwala wdrożyć wyrafinowane uczenie maszynowe programów na znacznie prostszym i tańszym sprzęcie. Wykonywanie analizy lub trenowanie sztucznej inteligencji na małych komputerach podłączonych do sieci nazywa się przetwarzaniem brzegowym lub, inaczej mówiąc, przetwarzaniem na brzegu sieci.

Dodanie sprzętowej sztucznej inteligencji do Raspberry Pi

Okazuje się, że wiele firm zaczęło budować wyspecjalizowane komputery AI, z których wiele można wykorzystać na Raspberry Pi. Zazwyczaj przekonasz się, że istnieją biblioteki lub opakowania Pythona, a często biblioteki Pythona TensorFlow, które obsługują korzystanie z tego. Do najciekawszych należą dwa

* Karta Intel Modvidius Neural Compute Stick (NCS): Karta Movidius NCS podłączana jest do portu USB Raspberry Pi lub innego komputera i zapewnia sprzętową obsługę analizy opartej na głębokim uczeniu się. Na przykład możesz użyć chmury Amazon do analizy, przetwarzania i klasyfikowania obrazów w chmurze z twojego małego systemu komputerowego, co przenosi twoje kosztowne obliczeniowo zadanie z Raspberry Pi do chmury. Kosztuje to pieniądze i przepustowość (i opóźnienie w twój system), aby to zrobić. Przeprowadzenie analizy za pomocą wytrenowanej sieci neuronowej głębokiego uczenia na krawędzi za pomocą pendrive'a NCS może pomóc i być może pozwolić na całkowite odłączenie urządzenia działającego na krawędzi sieci od Internetu. Działa około 60 razy szybciej niż analiza obrazu na Raspberry Pi i kosztuje mniej niż 100 USD. Możesz wykonywać rozpoznawanie twarzy, analizę tekstu, monitorowanie i konserwację za pomocą tego kija NCS. Całkiem fajne! Jest jednak jedna koncepcja, którą musimy tutaj podkreślić. Kij NCS służy do wykonywania analiz i wnioskowania na danych, ale nie do trenowania modeli! Nadal musisz zbudować i wyszkolić modele. Ma dobry interfejs z Keras i TensorFlow, więc można to zrobić w rozsądny sposób. Pomyśl o tym jak o akceleratorze do wykorzystania w końcowym projekcie po zakończeniu szkolenia.

* Akcelerator Google Edge TPU: Google Edge TPU (jednostka przetwarzająca tensor) ma gniazdo USB typu C, które można podłączyć do systemu opartego na systemie Linux, aby zapewnić przyspieszoną analizę uczenia maszynowego i wyciąganie wniosków. Czy słowo tensor brzmi znajomo? Tensory to macierze, jak w naszych przykładach sieci neuronowych. Cóż, okazuje się, podobnie jak powyższy kij Intel NCS, to urządzenie polega na wykonywaniu wyszkolonych modeli uczenia maszynowego. Nadal trenujemy sieci uczenia maszynowego przy użyciu innych technik, a następnie wykonujemy model na patyku.

KOMENTARZ KOŃCOWY DOTYCZĄCY AKCELERATORÓW UCZENIA MASZYNOWEGO

O chłopie. W ciągu najbliższych czterech lat tego typu wyspecjalizowany sprzęt do uruchamiania modeli uczenia maszynowego eksploduje. Zobaczysz wiele różnych architektur i rozwiązań pochodzących od Google, Intel, Nvidia, AMD, Qualcomm i wielu innych mniejszych firm z całego świata. Wszyscy zaczynają wspinać się na modę sprzętowego akceleratora AI

Sztuczna inteligencja w chmurze

W branży technologicznej wszyscy uwielbiają używać modnych słów, takich jak chmura. Często użycie takiego języka skutkuje arbitralnymi i niejasnymi terminami, które sprawiają, że konsumenci (a nawet wyrafinowani specjaliści techniczni) nie są pewni, co myślą, gdy firma mówi „twoje dane są w chmurze" lub „możesz pracować w chmurze", nie ma to nic wspólnego z byciem białym, puszystym lub nadziemnym. Twoje dane „w chmurze" znajdują się na ziemi i są przechowywane gdzieś w centrum danych z grupą serwerów, które są bardziej podobne do komputera PC lub Mac, niż myślisz. Niektórzy definiują chmurę jako oprogramowanie lub usługi działające w Internecie, a nie na komputerze lokalnym. Jest to do pewnego stopnia poprawne, ale tak naprawdę nic nie działa w Internecie; działa na komputerach podłączonych do Internetu. Zrozumienie, że oprogramowanie in-the-cloud działa na serwerach i nie jest „po prostu tam", bardzo szybko wyjaśnia chmurę i jej funkcje. Jeśli masz dwa komputery połączone w sieć i używasz drugiego komputera jako serwera danych, masz własną „chmurę". Dotyczy to podstawowych usług, takich jak przechowywanie danych w chmurze, ale w chmurze dostępne jest znacznie więcej niż tylko przechowywanie i tutaj robi się naprawdę interesująco. Zaletą korzystania z chmury jest to, że możesz korzystać z usług i pamięci masowej niedostępnych dla Ciebie w Twojej sieci lokalnej i (w jednym z najważniejszych zmieniaczy gry przetwarzania w chmurze) możesz dynamicznie zwiększać i zmniejszać wykorzystanie w zależności od potrzeb obliczeniowych. Korzystanie z chmury wymaga dostępu do Internetu. Niekoniecznie przez 100 procent czasu (możesz odpalić proces w chmurze, a potem wrócić do niego później), ale czasami

potrzebujesz połączeń. Ogranicza to chmurę w aplikacjach, takich jak samojezdne samochody, które nie gwarantują dobrego dostępu do Internetu przez cały czas. Co ciekawe, ten tryb „odpal i zapomnij" jest przydatny w przypadku urządzeń IOT (Internet of Things), w których nie chcesz pozostawać w kontakcie z siecią przez cały czas ze względów energetycznych. Jak więc korzystać z chmury? To zależy od usługi i dostawcy, ale w aplikacjach uczenia maszynowego najczęstszym sposobem jest skonfigurowanie Pythona na komputerze, który wywołuje funkcje i aplikacje oparte na chmurze. Wszyscy dostawcy chmury dostarczają przykładów. Jaki jest doskonały konsumencki przykład wykorzystania chmury? Amazon Echo i Alexa. Słucha cię, kompresuje dane mowy, wysyła je do chmury Amazon AWS, tłumaczy i interpretuje twoje dane, a następnie odsyła ustną odpowiedź lub polecenia, aby zapaliły się twoje światła. Istnieje wielu dostawców usług w chmurze do przechowywania danych i usług, a wciąż przybywa nowych. Czterej najwięksi dostawcy chmury dla sztucznej inteligencji to:

* Chmura Google

* Amazon Web Sevice

* Chmura BM

*Microsoft Azure

Chmura Google'a

Chmura Google jest prawdopodobnie najbardziej skoncentrowanym na sztucznej inteligencji dostawcą chmury. Możesz uzyskać dostęp do TPU (jednostek przetwarzania tensorów) w chmurze, które, podobnie jak powyższy kij Google TPU, mogą przyspieszyć Twoje aplikacje AI. Wiele funkcji chmury Google odzwierciedla podstawowy zestaw umiejętności firmy – wyszukiwanie. Na przykład Cloud Vision API może wykrywać obiekty, obiekty i punkty orientacyjne na obrazach. Kilku znakomitych studentów z University of Idaho tworzy aplikację Smart City o nazwie ParkMyRide, która wykorzystuje zasilany energią słoneczną aparat oparty na Raspberry Pi do robienia zdjęć ulicy i określania dostępności miejsc parkingowych przy użyciu Google Cloud Vision API. Oprogramowanie wysyła zdjęcie ulicy do Google i zwraca liczbę znalezionych samochodów oraz ich położenie na zdjęciu. Następnie dostarczają te informacje do aplikacji na smartfona, która wyświetla je graficznie. Całkiem schludnie. Inne polecane usługi w chmurze Google to: aplikacje do wyszukiwania treści wideo i pakiety zamiany mowy na tekst/tekstu na mowę (pomyśl Google Home – bardzo podobny do Amazon Alexa). Podobnie jak Amazon i Microsoft, Google używa własnych aplikacji opartych na sztucznej inteligencji do tworzenia nowych usług dla klientów.

Amazon Web Services

Amazon Web Services (AWS) koncentruje się na wykorzystywaniu wiedzy konsumenckiej w zakresie sztucznej inteligencji i dostarczaniu tej wiedzy firmom. Wiele z tych usług w chmurze jest opartych na wersjach produktów konsumenckich, więc na przykład wraz z ulepszeniami Alexa poprawiają się również usługi w chmurze. Amazon oferuje nie tylko tekst i język naturalny, ale także narzędzia do wizualizacji/tworzenia uczenia maszynowego, rozpoznawania wizji i analizy.

Chmura IBMa

W ciągu ostatnich kilku lat chmura IBM zyskała złą reputację za to, że jest trudna w użyciu. Jednym z głównych powodów było to, że było tak wiele różnych opcji na tak wielu różnych platformach, że prawie niemożliwe było ustalenie, od czego zacząć. W ciągu ostatnich kilku lat znacznie się poprawiło. IBM połączył swoje trzy duże dywizje (usługi chmurowe IBM BlueMix, usługi danych SoftLayer i grupa Watson AI) w jedną grupę pod marką Watson. Nadal dostępnych jest ponad 170 usług, więc nadal

trudno jest zacząć, ale istnieje znacznie lepsza kontrola i spójność procesu. Ich środowisko uczenia maszynowego nazywa się Watson Studio i służy do budowania i trenowania modeli AI w jednym zintegrowanym środowisku. Zapewniają również ogromne katalogi wiedzy z możliwością przeszukiwania i mają jedną z lepszych dostępnych platform zarządzania IOT (Internet of Things). Jedną z fajnych rzeczy, które mają, jest usługa o nazwie Watson Personality Insights, która przewiduje cechy osobowości, potrzeby i wartości za pomocą tekstu pisanego. Co Watson Personality pomyślałby o autorach tej książki? Przepuścimy tekst gotowej książki przez Watsona i poinformujemy Cię o tym na blogu Wiley.

Microsoft Azure

Microsoft Azure kładzie nacisk na programistów. Dzielą swoją ofertę AI na trzy kategorie AI:

* Usługi sztucznej inteligencji

* Narzędzia i frameworki AI

* Infrastruktura sztucznej inteligencji

Podobnie jak Amazon i Google, ich aplikacje AI są zbudowane na produktach konsumenckich wyprodukowanych przez Microsoft. Platforma Azure obsługuje również wyspecjalizowane układy FPGA (programowalne macierze bramek — pomyśl o sprzęcie, który można zmienić przez programowanie) i zbudowała infrastrukturę obsługującą szeroką gamę akceleratorów. Microsoft jest jednym z największych, jeśli nie największym klientem chipów Intel Movidius. Mają produkty do uczenia maszynowego, zestawy narzędzi IOT i usługi zarządzania, a także pełny i bogaty zestaw usług danych, w tym bazy danych, obsługę procesorów graficznych i niestandardową krzemową infrastrukturę AI, a także usługę kontenerową, która może przekształcić aplikacje wewnętrzne w aplikacje w chmurze. Platforma Microsoft Azure jest tą, na którą należy zwrócić uwagę na całkiem spektakularne innowacje.

AI na karcie graficznej

Karty graficzne są integralną częścią komputera PC od dziesięcioleci. Ludzie często szukają najnowszych i najlepszych kart graficznych, aby uczynić swoje komputery lepszymi maszynami do gier. Jedna rzecz staje się oczywista po pewnym czasie: chociaż szybkość procesora jest ważna, jakość i architektura karty graficznej ma większe znaczenie. Dlaczego? Ponieważ przetwarzanie grafiki o wysokiej rozdzielczości jest kosztowne obliczeniowo, a sposobem na rozwiązanie tego problemu jest zbudowanie kart graficznych z komputerów zaprojektowanych do wykonywania grafiki w celu podziału obciążenia. Tak narodził się GPU (jednostka przetwarzania grafiki), wyspecjalizowany rdzeń komputera przeznaczony do pracy z grafiką. Nvidia i inni zaczęli budować karty graficzne z wieloma procesorami graficznymi, co radykalnie poprawiło rozdzielczość wideo i liczbę klatek na sekundę w grach. Należy pamiętać, że algorytmy graficzne są konstruowane przy użyciu struktur danych zwanych macierzami (lub tensorami), które są przetwarzane w potokach. Czekać. Tensory? Matryce? Brzmi to podejrzanie jak struktury danych, których używamy w sztucznej inteligencji i uczeniu maszynowym. Ze względu na sposób, w jaki odbywa się i wdraża uczenie maszynowe i głębokie, procesory graficzne okazały się przydatne i skuteczne. Głębokie uczenie się opiera się na wielu różnych typach sieci neuronowych, a my trenujemy i używamy tych sieci za pomocą tensorów. Niezależnie od rodzaju zastosowanej sieci neuronowej, wszystkie techniki polegają na wykonywaniu złożonych operacji statystycznych. Podczas operacji szkoleniowych (uczących się) do sieci przesyłanych jest wiele obrazów lub punktów danych, a następnie trenowane z prawidłową klasyfikacją lub poprawną odpowiedzią. Korelujesz miliony tensorów (macierzy), aby zbudować model, który uzyska właściwy wynik. Aby przyspieszyć szkolenie,

operacje te można wykonywać równolegle, co okazuje się bardzo dobrym wykorzystaniem GPU na karcie graficznej. Pojedynczy rdzeń GPU jest znacznie prostszy niż rdzeń procesora, ponieważ jest przeznaczony do określonego, a nie ogólnego celu. To sprawia, że budowanie wielordzeniowych układów GPU jest tańsze niż budowanie wielordzeniowych układów procesorów. Rozpowszechnienie kart graficznych z wieloma rdzeniami GPU sprawiło, że komputery te idealnie nadają się do zastosowań związanych z uczeniem maszynowym. Połączenie potężnego procesora wielordzeniowego i wielu procesorów graficznych może radykalnie przyspieszyć programy uczenia maszynowego. W szczególności TensorFlow ma wersje oprogramowania przeznaczone do pracy z kartami GPU, usuwając wiele komplikacji związanych z korzystaniem z tych kart. Aby spojrzeć na to z perspektywy, nasz Raspberry Pi 3B ma 4 rdzenie procesora iw pewnym sensie 4 rdzenie GPU. Jedna z najnowszych kart graficznych Nvidii ma 3584 rdzenie. Za pomocą tych kart graficznych GPU o dużej liczbie rdzeni można przeprowadzać wiele szybkich szkoleń i uruchamiać sieci uczenia maszynowego. Płyty oparte na GPU nie są ostatnim krokiem w ewolucji wyspecjalizowanych komputerów i sprzętu do obsługi aplikacji AI. Zaczynają pojawiać się jeszcze bardziej wyspecjalizowane chipy. Według ostatnich obliczeń ponad 50 firm pracuje nad chipami, które przyspieszą funkcje sztucznej inteligencji. Kiedy wcześniej omawialiśmy ofertę chmury Microsoft Azure, wspomnieliśmy, że Microsoft zbudował infrastrukturę do obsługi sprzętu do akceleracji sztucznej inteligencji w chmurze. To jeden z głównych powodów, dla których warto obserwować, co robi Microsoft. Przyszłość należy do coraz bardziej wyspecjalizowanego sprzętu, zwłaszcza że wyspecjalizowany sprzęt staje się coraz łatwiejszy w obsłudze od strony oprogramowania użytkownika.