Dynamiczne sieci neuronowe dla efektywnego uczenia maszynowego

Znaczenie projektu

Nauka oraz przemysł w coraz większym stopniu wykorzystują metody uczenia maszynowego, w szczególności sztuczne sieci neuronowe. Jednocześnie modele te stają się bardziej skomplikowane i, aby przetwarzać eksponencjalnie rosnące ilości danych, wykorzystują coraz więcej zasobów obliczeniowych. Trend ten jest widoczny w wielu dziedzinach, począwszy od przetwarzania obrazów medycznych aż po robotykę i przemysł samochodowy.

Jednym z przykładów ogromnych kosztów obliczeniowych wykorzystania modeli uczenia maszynowego są eksperymenty fizyki wysokich energii, gdzie zbierane i przetwarzane są niespotykane dotychczas ilości danych. Przykładowo, eksperyment ALICE przy Wielkim Zderzaczu Hadronów w CERN, największym akceleratorze cząstek na świecie, rejestruje petabajty danych na godzinę. Dane te następnie przetwarzane są przez liczne modele sztucznej inteligencji. Złożone obliczenia wykonywane przez te modele skutkują długim czasem przetwarzania, wysokim poborem energii, a w konsekwencji znaczącym śladem węglowym pozostawionym przez infrastrukturę obliczeniową.

Cel projektu

W odpowiedzi na wymienione powyżej problemy, w naszym projekcie proponujemy wykorzystać **dynamiczne sieci neuronowe**, czyli modele zdolne do adaptacji połączeń występujących w sieci w zależności od dostępnych zasobów obliczeniowych. Modele te mogą na przykład wykorzystywać obliczenia wykonane we wcześniejszych krokach, co pozwala ograniczyć czas przetwarzania oraz zużycie pamięci. Dynamiczne sieci neuronowe łączą zagadnienia takie jak *obliczenia warunkowe, reużywalne komponenty, inferencja z wykorzystaniem częściowej informacji* oraz *uczenie ciągłe*.

W porównaniu do standardowych metod optymalizacji sieci neuronowych, zamiast ograniczać liczbę obliczeń czy pamięć wykorzystywaną przez modele, skupiamy się na dynamicznej adaptacji struktury oraz obliczeń wykonywanych w sztucznych sieciach neuronowych do dostępnych zasobów, takich jak:

- obliczenia wykonane podczas poprzednich etapów przetwarzania,
- częściowa informacja dostępna w czasie działania modelu,
- moduły architektoniczne zaprojektowane w modelach uczenia ciągłego.

W szczególności, skupiamy się na projektowaniu modeli zdolnych nie tylko rozwiązywać powierzone im zadania, ale również uczyć się, w jaki sposób robić to jak najefektywniej. Stawiamy hipotezę badawczą, że dynamiczna adaptacja komponentów składowych sieci neuronowych oraz ścieżek przetwarzania informacji w sieciach może znacząco zwiększyć wydajność tych modeli. Motywowani tym założeniem, w projekcie planujemy rozwijać modele oparte o **dynamiczne sieci neuronowe**, które skupiają się na efektywności obliczeń oraz redukcji wykorzystania zasobów. Celem projektu jest więc stworzenie wydajnych modeli uczenia maszynowego zdolnych do dynamicznej adaptacji w zależności od dostępnych zasobów obliczeniowych oraz ich ewaluacja w jednym z największych eksperymentów stworzonych przez człowieka, Wielkim Zderzaczu Hadronów w CERNie.

Oczekiwane rezultaty

Głównym rezultatem naszego projektu będzie zestaw nowatorskich modeli uczenia maszynowego zdolnych do dynamicznej adaptacji ścieżek przetwarzania informacji oraz efektywnie wykorzystujących komponenty, z których się składają. Modele te będą dostosowywać swoje działanie do dostępnych dla nich zasobów, takich jak częściowa informacja lub obliczenia wykonywane w poprzednich krokach przetwarzania.

Zamierzamy testować opracowane metody w praktyce, wdrażając je w aplikacjach związanych z eksperymentami fizyki wysokich energii. Środowisko to jest idealną przestrzenią do naszych badań ze względu na praktyczne i rzeczywiste ograniczenia występujące przy przepływie ogromnych ilości danych zbieranych podczas zderzeń cząstek.

Praktycznym rezultatem projektu będzie integracja opracowanych wydajnych, dynamicznych sieci neuronowych w infrastrukturze IT będącej częścią największego eksperymentu fizyki wysokich energii w CERN.