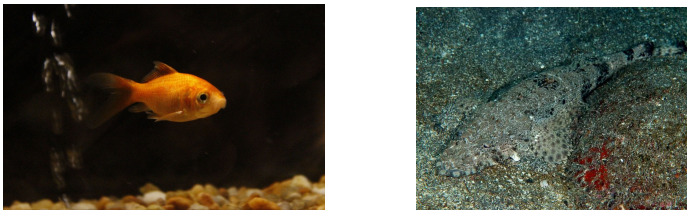
Wydajne obliczeniowo dynamiczne sieci neuronowe

Streszczenie popularnonaukowe

W ostatniej dekadzie w głębokim uczeniu nastąpił masowy wzrost zainteresowania naukowców i przemysłu. Na ten rozwój złożyły się zarówno postępy w rozwoju algorytmów i w projektowaniu modeli uczenia maszynowego, jak i rosnąca z czasem moc obliczeniowa. Zwyczajne skalowanie architektury modelu często skutkuje poprawionymi wynikami. W efektcie mamy stały wzrost średniego rozmiaru parametrów i narzutu obliczeniowego najnowszych modeli. Koszt urządzeń i energii wymagany do wytrenowania lub ewaluacji takich modeli jest często poza zasięgiem możliwości finansowych większości instytucji badawczych czy małych firm, ograniczając ich zdolności do rywalizowania z większymi podmiotami.



Rysunek 1: Dwie próbki tej samej klasy o znacznie różniącym się poziomie trudności.

Celem tego projektu jest zbadanie oraz rozwój metod, które skupiają się na trenowaniu modeli wydajnych obliczeniowo. Projekt skupi się na dynamicznych sieciach neuronowych, czyli na grupie modeli, w których graf obliczeniowy jest warunkowany aktualnym przykładem. Przykładowa para wejść do modelu jest przedstawiona na Rysunku 1. Pierwszy obrazek jest łatwiejszy do sklasyfikowania, więc prawdopodobnie do uzyskania prawidłowej odpowiedzi potrzebuje mniej obliczeń niż drugi obrazek. Opisany sposób oszczędzania obliczeń odzwierciedla funkcjonowanie ludzkiego mózgu.

Istniejące modele dynamiczne zazwyczaj skupiają się tylko na jednym sposobie warunkowania obliczeń. Sieci ze wczesnymi wyjściami polegają na dołączeniu dodatkowych głów klasyfikacyjnych do pośrednich warstw modelu, z których korzystają podczas inferencji poprzez zwrócenie wczesnej odpowiedzi. Mikstury ekspertów wybierają stosowne moduły ekspertów, do których przekazane zostanie wejście. Inna grupa modeli dostosowuje ilość obliczeń do każdej części wejścia osobno, np. rejonów obrazka. O ile niektóre z tych modeli znacząco zmniejszają wymagania obliczeniowe przy zachowaniu oryginalnych wyników, to związek pomiędzy tymi metodami nigdy nie został dogłębnie zbadany. Ponadto, mikstury ekspertów były zazwyczaj używane do utrzymania kosztu obliczeń przy jednoczesnym skalowaniu liczby parametrów w górę, a więc do powiązanego, ale innego celu niż samo redukowanie obliczeń.

W tym projekcie zamierzamy przeprowadzić serię analiz, eksperymentów i ewaluacji zaproponowanych modeli, które uzupełnią wspomniane luki. Po pierwsze, projektujemy nowatorską metodę inspirowaną metodami *boostingu*, która dostosowuje ilość obliczeń dla każdego symbolu niezależnie. To pozwoli nie tylko na redukcję narzutu obliczeniowego, ale również na poszeregowanie przykładów według wybranych dla nich ilości obliczeń. Intuicyjnie to poszeregowanie pokaże które przykłady są dla danej metody łatwe, oraz które są trudne. Po sprawdzeniu podobnych poszeregowań dla innych typów metod możemy zadać następujące pytanie: czy różne typy sieci dynamicznych szeregują przykłady podobnie? Zakładając aktywacje typu *Rectified Linear Unit* możemy również postawić pytanie: jaka część aktywacji w warstwach pośrednich modeli statycznych jest zerowa? Odpowiedź na te pytania może prowadzić do wydajniejszych obliczeniowo modeli, a więc bardziej dostępnego środowiska naukowego, dłuższego czasu życia baterii oraz ograniczenia emisji gazów cieplarnianych.