Nr rejestracyjny: 2022/45/N/ST6/02222; Kierownik projektu: mgr Sebastian Dariusz Jaszczur

# Warunkowe obliczenia w głębokich sieciach neuronowych

Głębokie sieci neuronowe są coraz częściej używane w wielu zastosowaniach. Ich użycie w dziedzinach takich jak rozpoznawanie obrazu oraz automatyczne tłumaczenie jest ważne dla wielu osób, ale niestety takie modele wymagają znaczących nakładów obliczeniowych zarówno w czasie treningu, jak i finalnego użycia. To wynika głównie z olbrzymiego rozmiaru tych modeli oraz używanymi obecnie architekturami.

Główną hipotezą projektu jest nieefektywne użycie mocy obliczeniowej przez obecnie stosowane sieci neuronowe. Co więcej, prawdopodobnie możnaby poprawić obecne architektury w różnych metrykach (takich jak czas treningu, czas ewaluacji, jakość predykcji modelu lub innych) za pomocą dynamicznych, warunkowych obliczeń w sieci neuronowej. Takie poprawki prowadziłyby do ulepszenia predykcji modelu z zachowaniem danego budżetu obliczeniowego, lub szybsze działanie modelu przy takiej samej jakości predykcji.

Możemy zastanowić się nad intuicyjnym wytłumaczeniem tej hipotezy, wyobrażając sobie konkretne zadanie dla sieci neuronowej. Załóżmy, że chcemy wytrenować model do rozpoznawania kotów i psów na zdjęciach, a także do rozpoznawania ich konkretnej rasy. Możemy wydzielić trzy podproblemy w tym zadaniu: (1) czy na obrazku jest kot czy pies, (2) jakiej rasy jest kot na obrazku, (3) jakiej rasy jest pies na obrazku. Możemy zauważyć, że podproblem (1) jest intuicyjnie łatwiejszy od pozostałych dwóch. Co więcej, jeśli dana sieć neuronowa zna już odpowiedź na podproblem (1) to, oczywiście, nie potrzebuje obliczać odpowiedzi dla obu zadań (2) oraz (3) - w zależności od odpowiedzi na podproblem (1) tylko jedno z tych zadań ma znaczenie, a drugie może być zignorowane.

Niestety, obecne architektury sieci neuronowych nie są w stanie warunkowo pominąć żadnych obliczeń. Będą więc liczyć odpowiedź na podproblem (3) nawet jeśli już wiedzą, że na zdjęciu nie ma żadnego psa. W tym projekcie będziemy rozpatrywać sposoby na umożliwienie sieciom neuronowym takiego pominięcia obliczeń oraz wytrenowanie ich do efektywnego używania takiej opcji.

Badania w tym projekcie mogą prowadzić do szybszych sieci neuronowych, z możliwością używania ich na tańszym sprzęcie, bez negatywnego wpływu na jakość ich predykcji. W rezultacie, zarówno badania jak i praktyczne użycie sieci neuronowych, i przetwarzania języka naturalnego w szczególności, byłyby bardziej dostępne dla każdego.