

Problem: Dostrajanie modeli językowych jest poza zasięgiem wielu zespołów

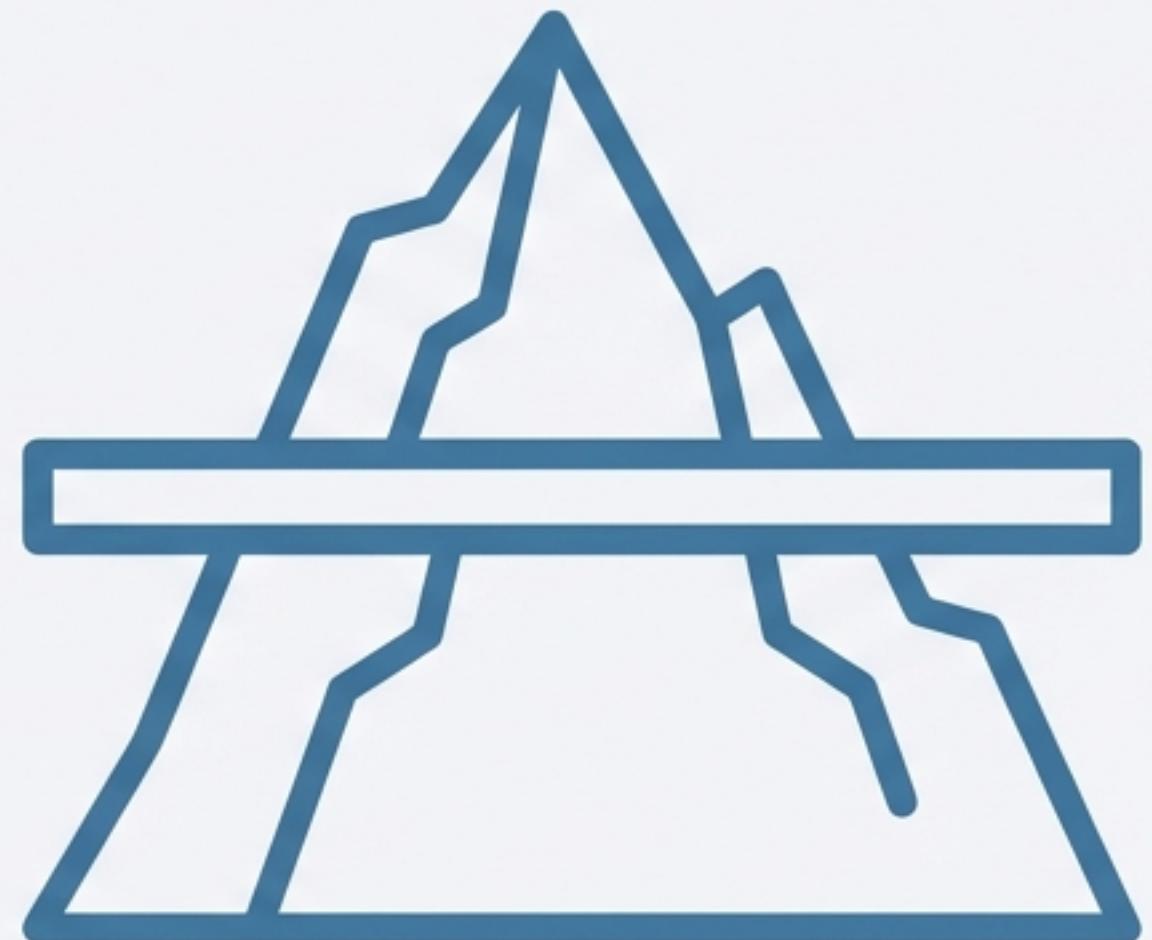
Pełne dostrajanie modelu LLaMA 65B w 16-bitowej precyzyji wymaga ponad **780 GB** pamięci GPU.

Tworzy to barierę nie do pokonania dla uniwersytetów, startupów i mniejszych zespołów badawczych.

Najpotężniejsze techniki dostrajania były dotąd zarezerwowane dla największych graczy w branży.

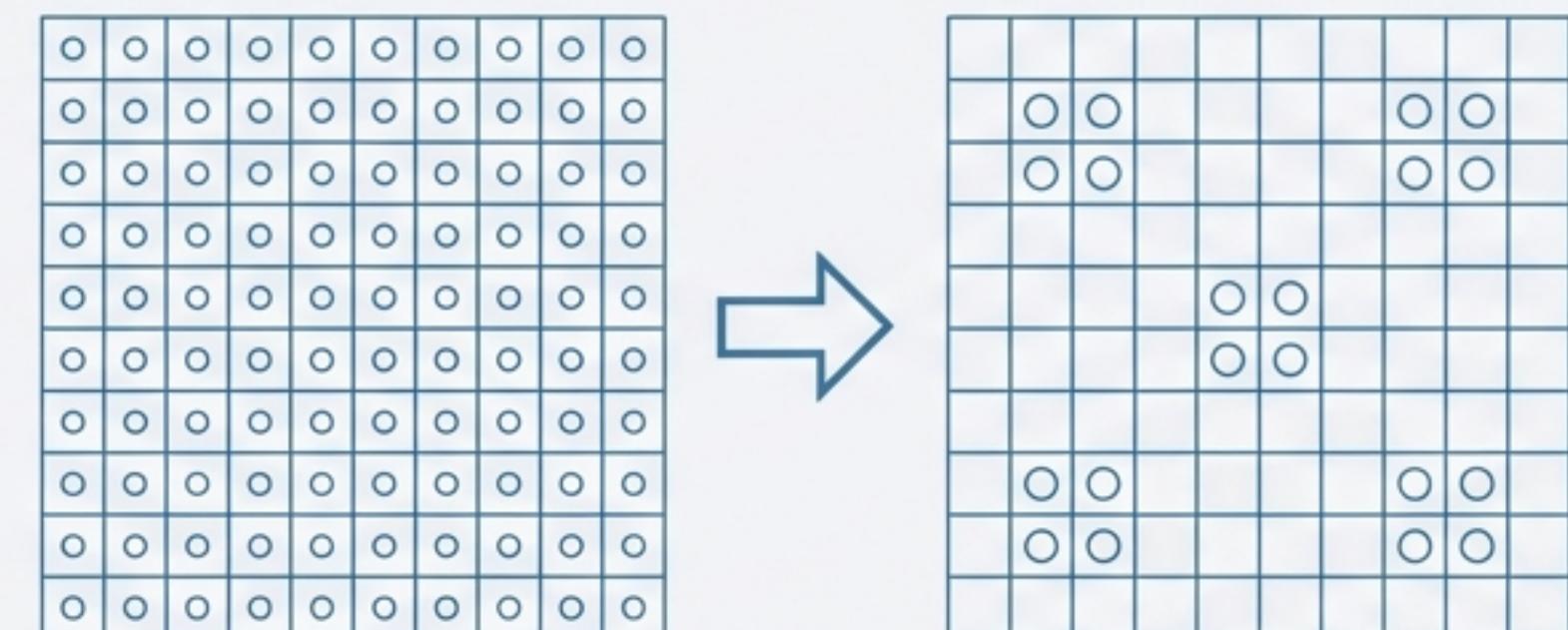
QLoRA obiecuje tę samą wydajność przy ułamku kosztów.

Umożliwia dostrojenie modelu 65B na pojedynczej, konsumenckiej karcie GPU o pojemności **48 GB**.



Koncepcja 1: Kwantyzacja - Zmniejszanie precyzji w celu oszczędności pamięci

- Kwantyzacja to proces redukcji precyzji numerycznej wag modelu w celu zmniejszenia jego rozmiaru.
- **Analogia:** Kompresja zdjęcia o wysokiej rozdzielczości z milionami kolorów do kilkuset kluczowych kolorów.
- Standardowe podejście: konwersja wartości 16-bitowych (FP16/BF16) na prostsze, np. 4-bitowe.
- Cel: Drastyczne zmniejszenie zapotrzebowania na pamięć przy zachowaniu zdolności modelu.
- Wagi w sieciach neuronowych zazwyczaj mają rozkład normalny (Gaussa), co jest kluczowe dla optymalizacji tego procesu.



16-bit

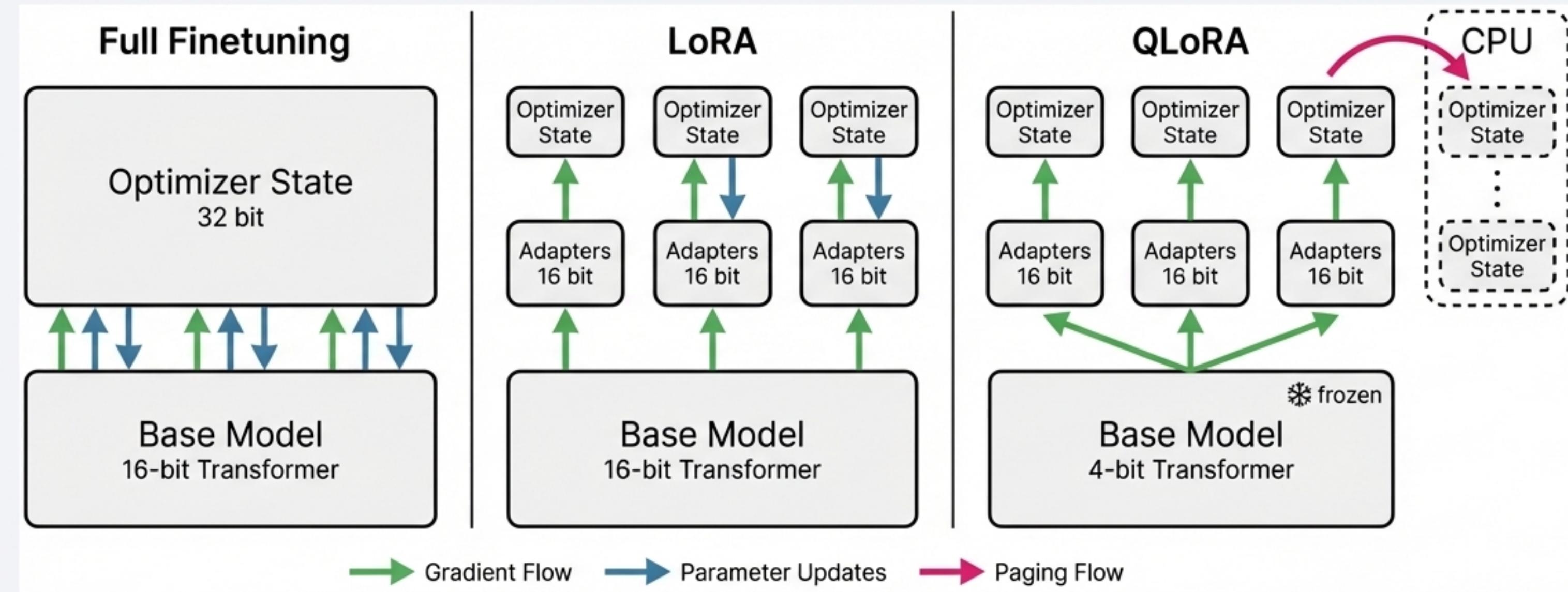
4-bit

Koncepcja 2: LoRA - Dostrajanie tylko niewielkiej części parametrów

- Zamiast modyfikować wszystkie 65 miliardów parametrów, zamrażamy wagi bazowego modelu.
- Trenujemy jedynie mały, dodatkowy zestaw parametrów nazywany "adapterami" (Low-Rank Adapters).
- **Analogia:** Zamiast remontować cały pałac, dobudowujemy mały, funkcjonalny pawilon.
- Główny model pozostaje w trybie "tylko do odczytu" podczas treningu. Cała nauka odbywa się w lekkich warstwach adapterów.
- Problem: Standardowe LoRA nadal wymaga przechowywania 16-bitowego modelu bazowego w pamięci GPU.



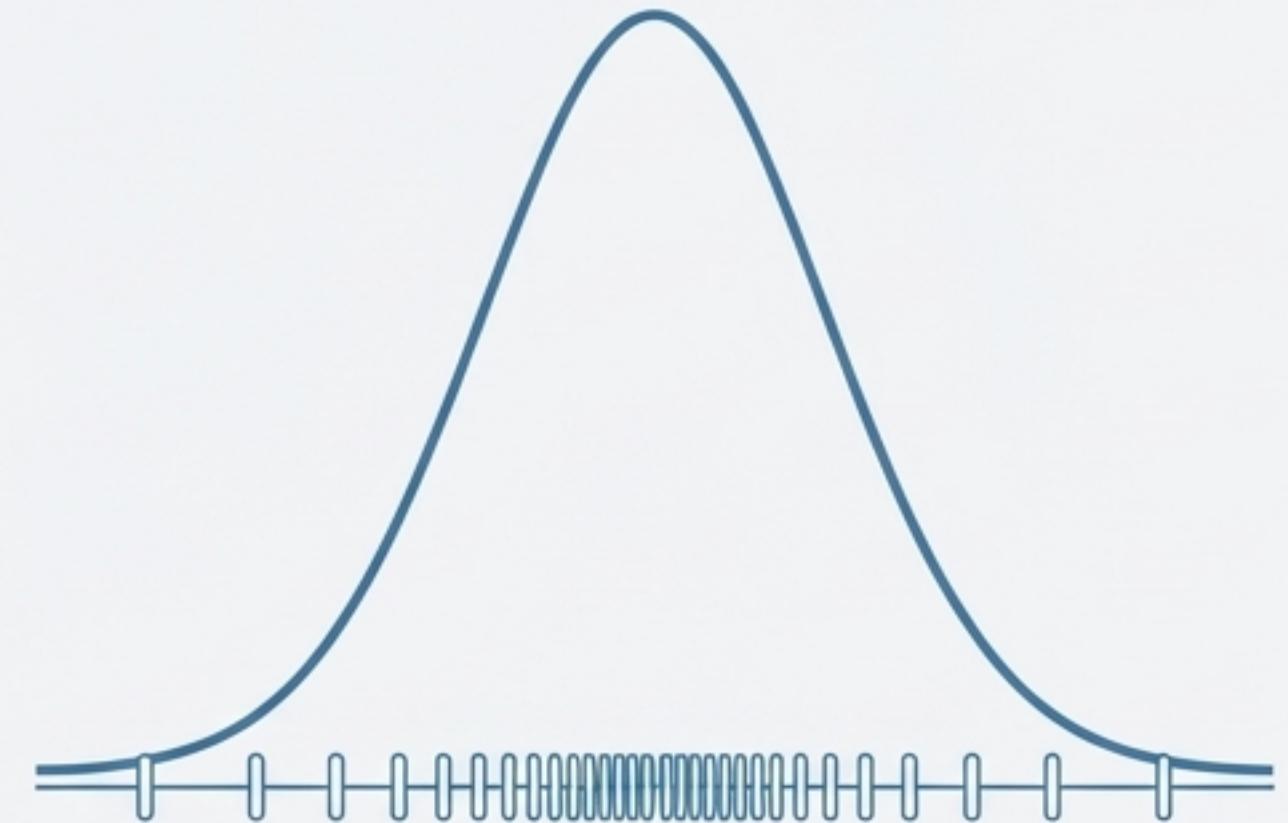
Przełom QLoRA: Połączenie kwantyzacji z adapterami LoRA



- QLoRA łączy zamrożony, 4-bitowy skwantyzowany model bazowy z trenowalnymi adapterami LoRA.
- Gradienty są propagowane wstecznie przez 4-bitowe wagi do adapterów, które są jedynymi aktualizowanymi parametrami.
- Dzięki temu redukcja pamięci dla modelu 65B jest ogromna: z >780 GB do <48 GB.
- Ten sukces jest możliwy dzięki trzem kluczowym innowacjom, które eliminują spadek wydajności.

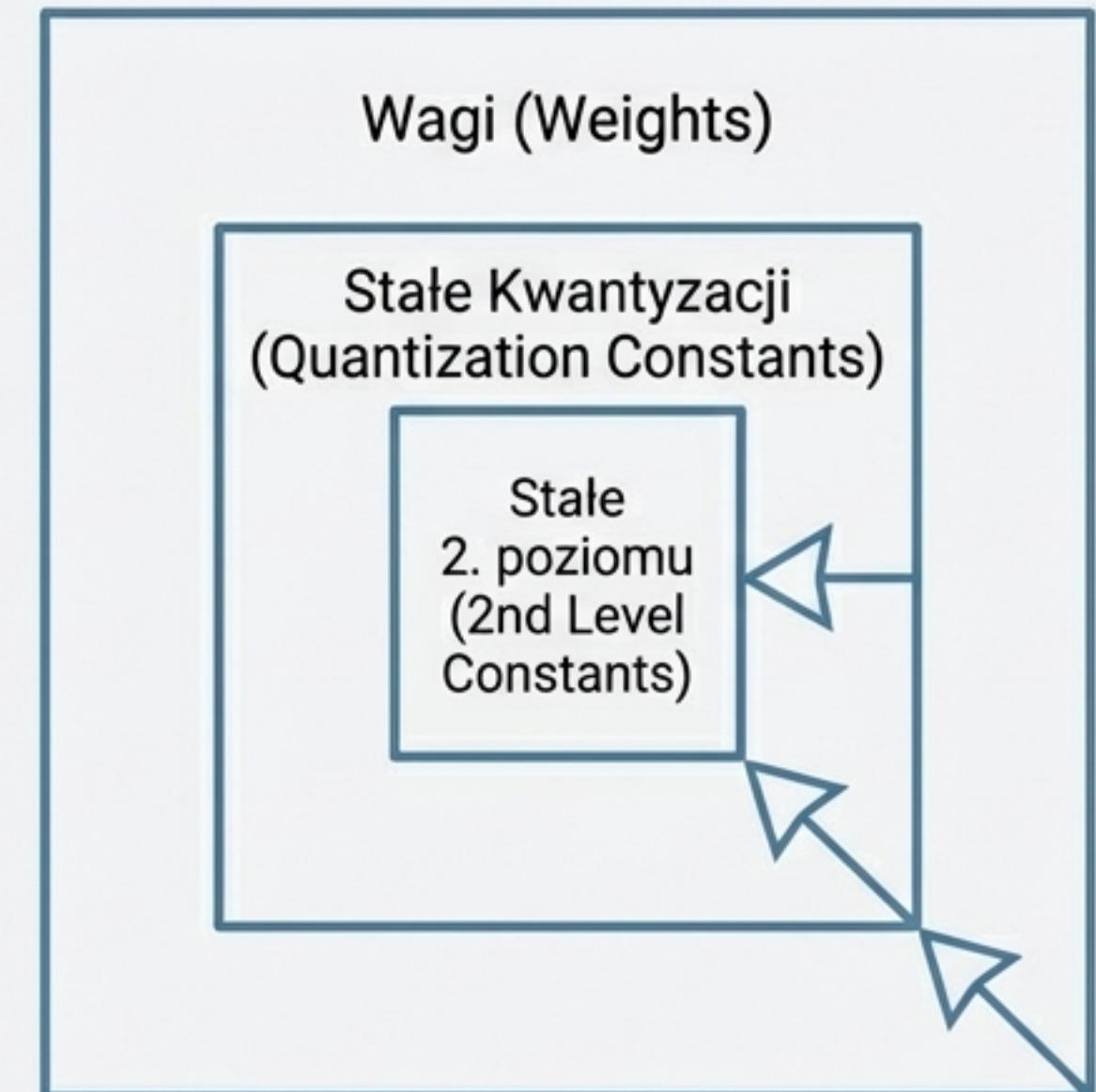
Innowacja 1: 4-bitowy NormalFloat (NF4) - Typ danych "szyty na miarę"

- Stworzono nowy typ danych, NF4, specjalnie dla wag sieci neuronowych o rozkładzie normalnym.
- Jest to typ "teoretycznie optymalny pod względem informacyjnym" dla takich danych.
- Zapewnia więcej poziomów kwantyzacji blisko zera i mniej na ekstremach, co idealnie pasuje do rozkładu wag.
- Minimalizuje błąd kwantyzacji w porównaniu do standardowych typów 4-bitowych (FP4). Badania pokazują, że QLoRA z FP4 osiąga o ~1 punkt procentowy gorsze wyniki na MMLU.



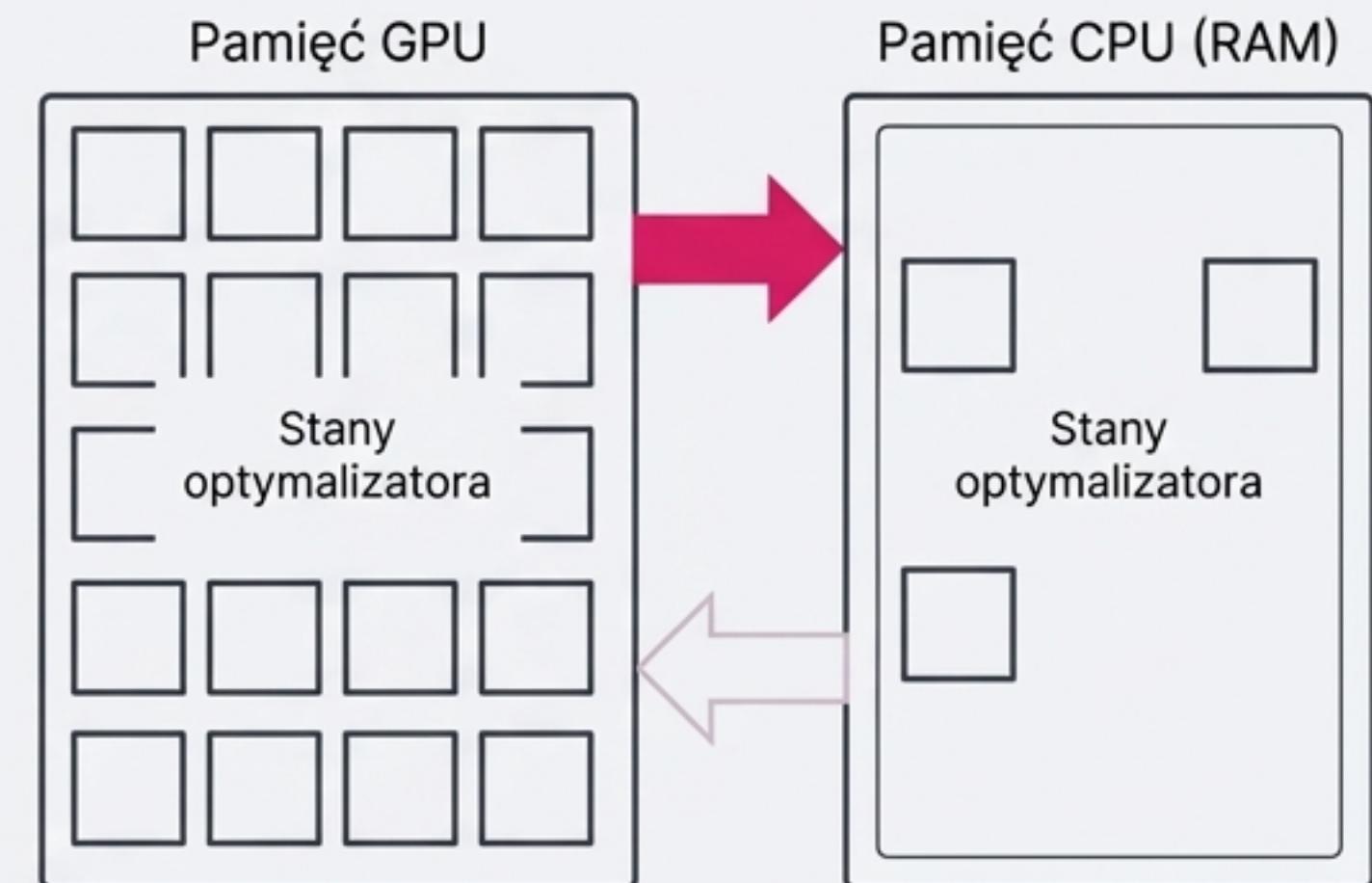
Innowacja 2: Podwójna Kwantyzacja - Kompresowanie danych o kompresji

- Sama kwantyzacja wymaga przechowywania metadanych – 'stałych kwantyzacji' dla każdego bloku wag.
- Podwójna kwantyzacja (Double Quantization, DQ) to proces kwantyzowania samych stałych kwantyzacji.
- "Incepcja optymalizacji": kompresujemy dane używane do kompresji.
- Oszczędza to średnio 0.37 bitów na parametr.
- Dla modelu 65B przekłada się to na oszczędność około **3 GB** pamięci – co może być różnicą między sukcesem a błędem braku pamięci.



Innowacja 3: Stronicowane Optymalizatory - Ochrona przed skokami pamięci

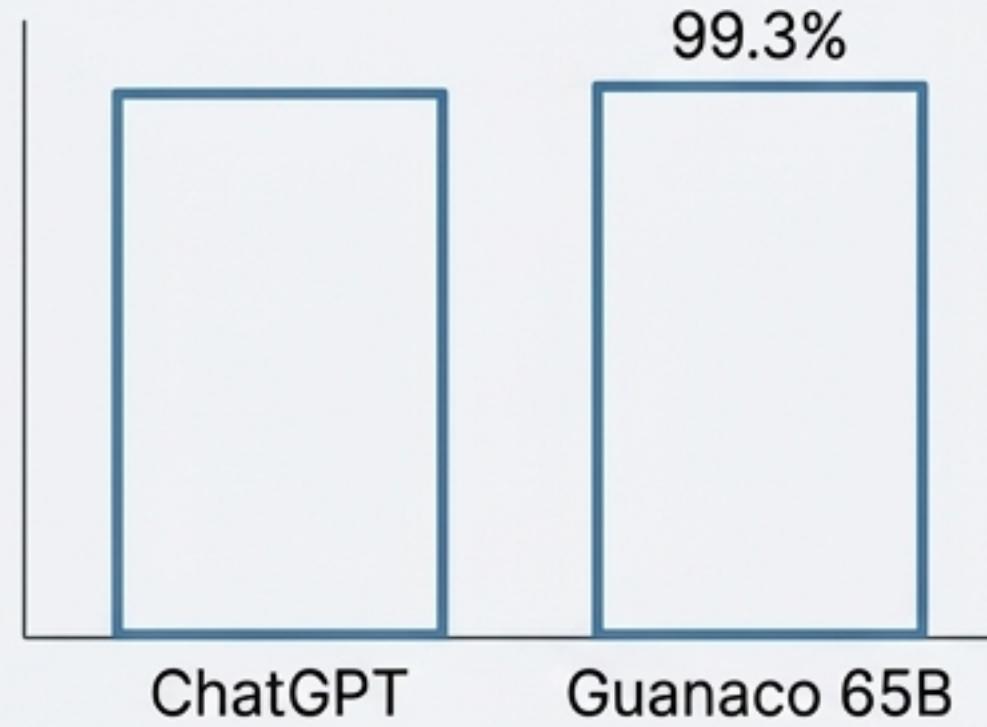
- Trening, zwłaszcza z długimi sekwencjami, może powodować nagłe skoki zużycia pamięci, prowadzące do błędów 'CUDA out of memory'.
- Stronicowane optymalizatory (Paged Optimizers) wykorzystują zunifikowaną pamięć NVIDIA.
- Gdy pamięć GPU się kończy, system automatycznie przenosi stany optymalizatora do pamięci RAM procesora.
- Gdy dane są ponownie potrzebne, są ładowane z powrotem do GPU.
- Zapewnia to stabilność treningu bez awarii, kosztem niewielkiego spowolnienia w rzadkich przypadkach.



Wyniki: QLoRA dorównuje wydajnością pełnemu dostrajaniu 16-bit

- **Kluczowy wniosek:** QLoRA z NF4 osiąga wydajność statystycznie **nieodróżnialną** od 16-bitowego dostrajania.
- Potwierdzone na benchmarkach GLUE, Super-NaturalInstructions i MMLU dla modeli RoBERTa, T5 i LLaMA.
- **Rodzina modeli Guanaco:** Najlepsze w momencie publikacji otwarte modele, stworzone przy użyciu QLoRA.
- Guanaco 65B osiąga **99.3%** wydajności ChatGPT na **benchmarku Vicuna**, trenowany w **24h** na jednym GPU.
- Guanaco 7B (wymaga 5GB pamięci) przewyższa wydajnością Alpaca 13B (wymaga 26GB) o ponad 20 punktów procentowych na tym samym benchmarku.

Wydajność na benchmarku Vicuna



Wydajność vs. Rozmiar





Głębsze wnioski: Jakość danych jest ważniejsza niż ich ilość

- **Efektywność QLoRA** umożliwiła przeprowadzenie ponad 1000 eksperymentów na dużą skalę, prowadząc do kluczowych obserwacji.
- **Jakość danych >> Ilość danych:** Zbiór OASST1 (9 tys. próbek) dał lepsze rezultaty w zadaniach konwersacyjnych niż FLAN v2 (450 tys. próbek).
- **Specjalizacja ma znaczenie:** Wysoka wydajność w jednym zadaniu (np. rozumienie tekstów akademickich w MMLU) nie gwarantuje wysokiej wydajności w innym (np. swobodna konwersacja w Vicuna).
- Nie istnieje uniwersalne podejście do dostrajania – dane treningowe muszą być precyzyjnie dopasowane do docelowego zastosowania.

Wpływ, Zastosowania i Ograniczenia Techniki QLoRA



Wpływ i Zastosowania

- **Demokratyzacja:** Zaawansowane dostrajanie staje się dostępne dla naukowców, startupów i hobbystów, zamykając lukę zasobów między nimi a korporacjami.
- **Prywatność:** Umożliwia personalizację modeli bez wysyłania danych na zewnętrzne serwery.
- **Zastosowania mobilne:** Otwiera drogę do dostrajania modeli na urządzeniach końcowych (laptopy, a w przyszłości smartfony).

Znane Ograniczenia i Słabości (modeli Guanaco)

- **Rozumowanie matematyczne:** Podatność na błędy, np. przy faktoryzacji liczb (potrafi podać dwie różne, błędne odpowiedzi w jednym zdaniu).
- **Podatność na 'prompt injection':** Model można łatwo oszukać w celu ujawnienia 'sekretnych' informacji.
- Autorzy przyznają: Brak bezpośredniego porównania z 16-bit dla modeli 33B/65B z powodu kosztów (wyniki ekstrapolowane z mniejszych modeli).

Skoro dostrajanie może w pełni odzyskać wydajność utraconą podczas agresywnej 4-bitowej kwantyzacji...



...jak daleko możemy się posunąć?

Czy modele skompresowane do 3 bitów, 2 bitów, a nawet wartości binarnych, również mogą osiągnąć pełną wydajność po dostrojeniu z użyciem adapterów?