deepsense.ai
BIG DATA SCIENCE

Large Language Models in limited hardware environments

April 17, 2023

Adam Maciaszek
Szymon Janowski

Agenda

1. Wstęp do wielkich modeli językowych (LLMs)
2. OpenAI vs. otwarte modele językowe
3. Aktualny research w obszarze modeli językowych
4. Metody optymalizacji wielkich modeli językowych
    a. Parallelism
    b. Mixed Precision Training
    c. ZeRO
    d. Kwantyzacja / destylacja / LoRA
5. Narzędzia do optymalizacji modeli językowych

We are data science experts

delivering AI-driven competitive edge for global leaders across industries

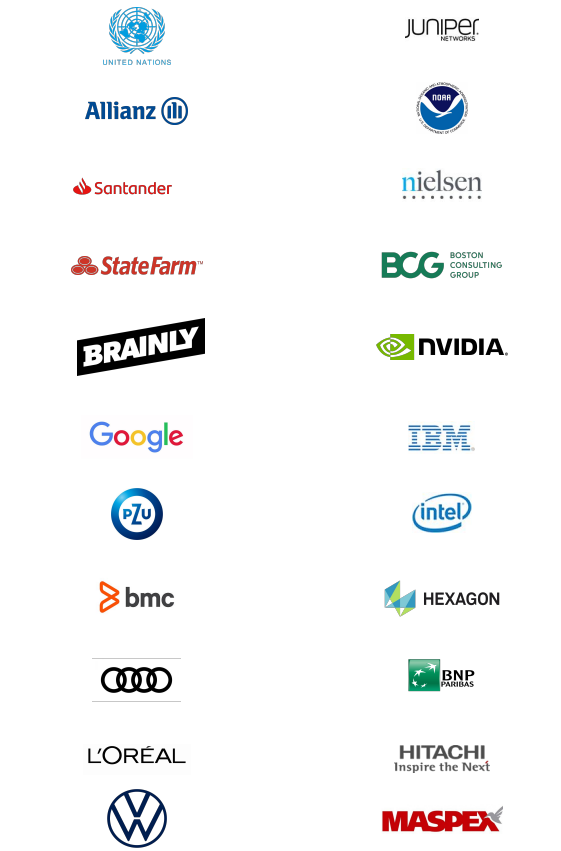
>150 **commercial AI projects** in the US and Europe.

>20 **R&D projects, for example with Intel and Google Brain** on topics involving reinforcement learning and generative models.

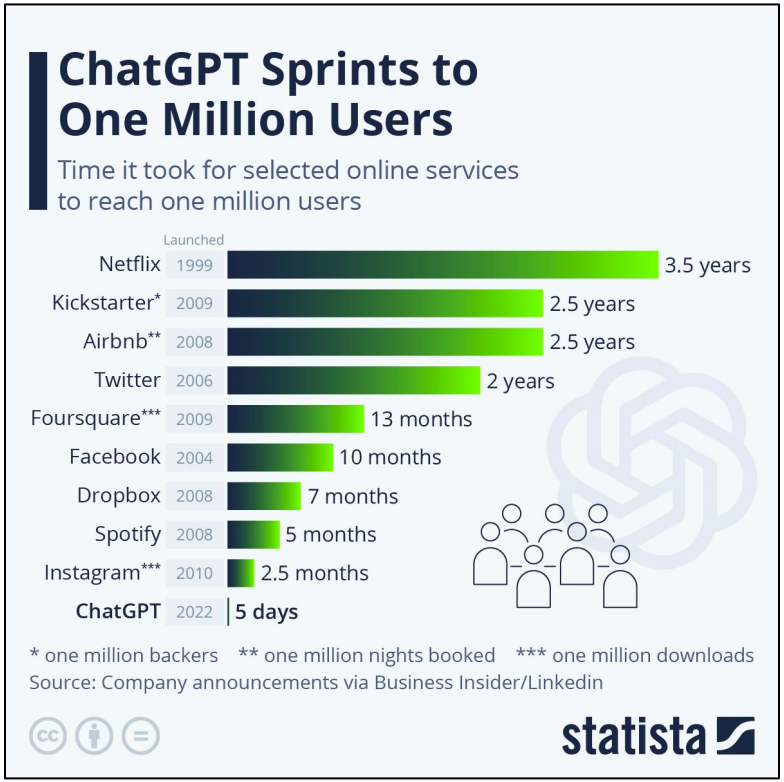
A winning team of over **120 world-class AI/ML experts** - data scientists and data engineers supported by software engineers.

Top-notch capabilities in: predictive analytics, computer vision and natural language processing.

deepsense.ai
BIG DATA SCIENCE



ChatGPT: hype jak nigdy



source: https://www.statista.com/chart/29174/time-to-one-million-users

GPT-4: Spark of AGI?

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

source: https://arxiv.org/abs/2303.12712

GPT-4: Spark of AGI?

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

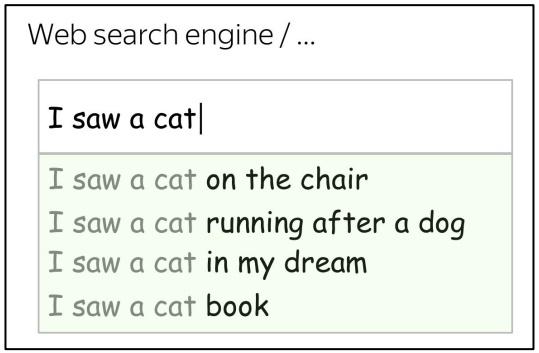
Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

source: https://arxiv.org/abs/2303.12712

Wstęp do modelowania językowego

Model językowy (ang. Language Model - LM) to algorytm przewidujący kolejny element w sekwencji wyrazów / tokenów:

- bazując na wielkich zbiorach danych tekstowych, uczy się prawdopodobieństwa poszczególnych sekwencji
- może być używany do generowania tekstów, które mogłyby być napisane przez człowieka
- jego prostą aplikacją jest autouzupełnianie w wyszukiwarkach i aplikacjach do wysyłania wiadomości



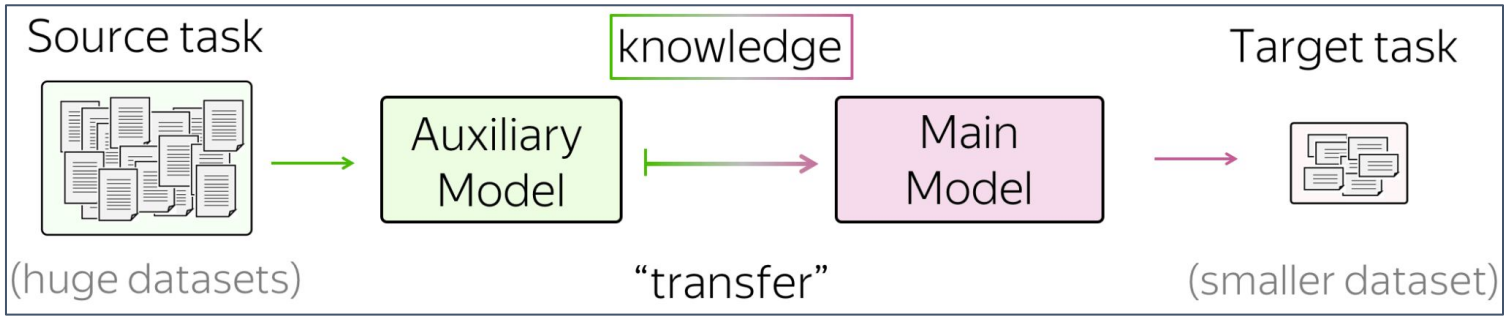
source: https://lena-voita.github.io/nlp_course/language_modeling.html

Transfer learning

Często stosowane podejście w użyciu modeli językowych, składa się z dwóch etapów:

- self-supervised pre-training - na wielkich zbiorach danych tekstowych, by nauczyć model "języka" i zależności pomiędzy słowami / tokenami
- supervised fine-tuning - adaptacja pretrenowanego modelu do konkretnego zadania przy użyciu danych z adnotacjami / etykietami

Zbiory danych w trakcie pretrenowania to mieszanka danych zebranych z Internetu, Wikipedii, książek, repozytoriów z kodem, itp.



source: https://lena-voita.github.io/nlp_course/transfer_learning.html

Rodzaje wielkich modeli językowych

There exist three types of LLMs:

- **dekodery** → generowanie tekstu (autouzupełnianie)

  GPT (2,3,4), ChatGPT, BLOOM, OPT, LLaMA

- **enkodery** → klasyfikacja sekwencji (klasyfikacja tekstu), klasyfikacja tokenów (rozpoznawanie encji - named entity recognition)

  BERT, RoBERTa, XLM-RoBERTa, HerBERT, TrelBERT

- **modele enkoder-dekoder** (seq2seq) → zadania text-to-text (tłumaczenie maszynowe, podsumowywanie tekstu, itp.)

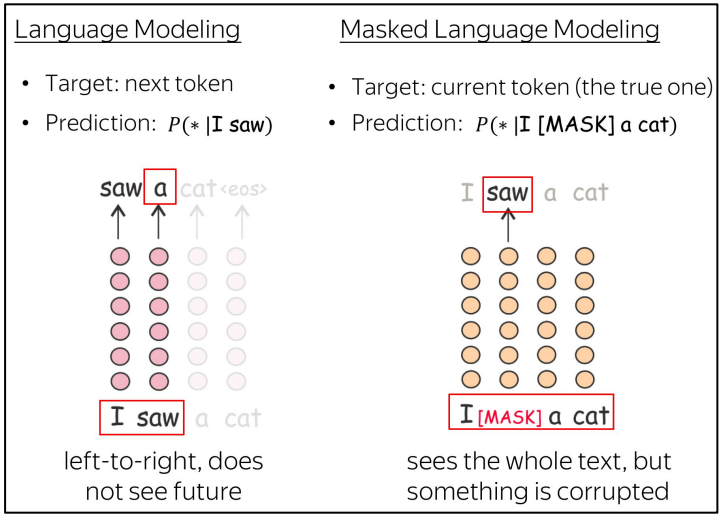
  T5, T0, Flan-T5

Decoders vs. encoders - pre-training

Dekodery: modelowanie językowe **Enkodery**: maskowane modelowanie językowe (MLM)

Enkodery wciąż są szeroko używane do zadań klasyfikacyjnych, ale nie są zdolne do generowania tekstu.

Zdecydowana większość ostatnio ogłoszonych modeli to dekodery.



source: https://lena-voita.github.io/nlp_course/transfer_learning.html

Użycie wielkich modeli językowych w biznesie?

Użycie wielkich modeli językowych w biznesie?

- OpenAI API (możliwe również za pomocą Azure OpenAI Service)
  - zbiór potężnych modeli
  - bardzo łatwe w użyciu
  - cena zależy od liczby zapytań i długości tekstów

Użycie wielkich modeli językowych w biznesie?

- OpenAI API (możliwe również za pomocą Azure OpenAI Service)
  - zbiór potężnych modeli
  - bardzo łatwe w użyciu
  - cena zależy od liczby zapytań i długości tekstów

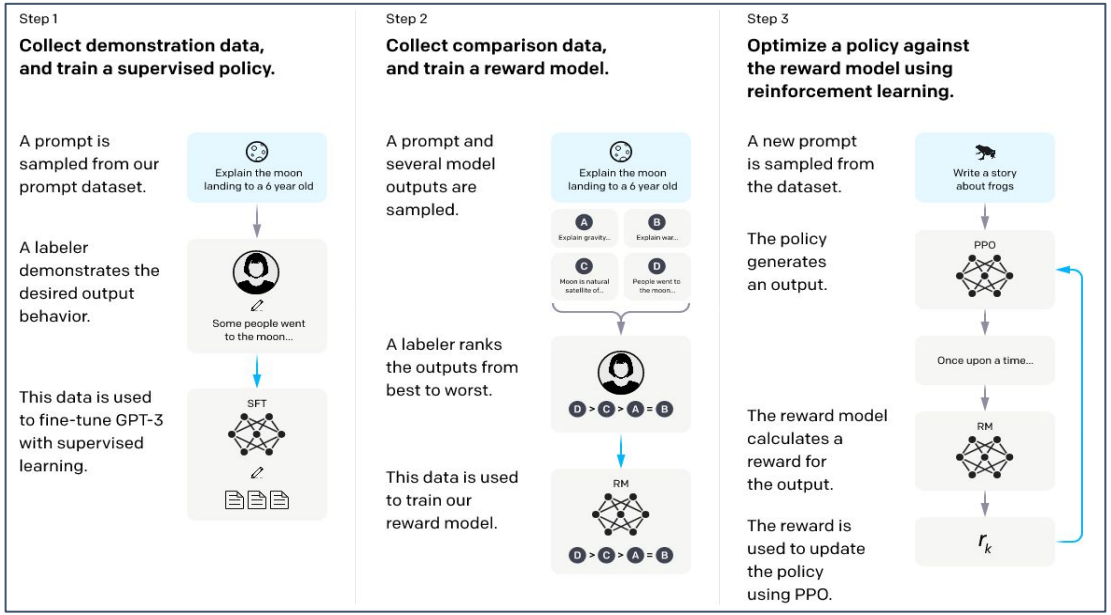
- poszukiwanie ogólnodostępnych alternatyw (np. modeli dostępnych w repozytorium HuggingFace)
  - możliwość dostosowania do indywidualnych potrzeb firmy / klienta, specyficznych danych
  - może działać na serwerach firmowych lub w chmurze (wymagane karty graficzne)

InstructGPT & ChatGPT

Training language models to follow instructions with human feedback (Ouyang et al., 2022)

W pierwszej połowie 2022, OpenAI opublikował model **InstructGPT**, znany również jako text-davinci-003.

Był on dotrenowywany na zbiorach instrukcji oraz przy użyciu techniki RLHF (Reinforcement Learning with Human Feedback).



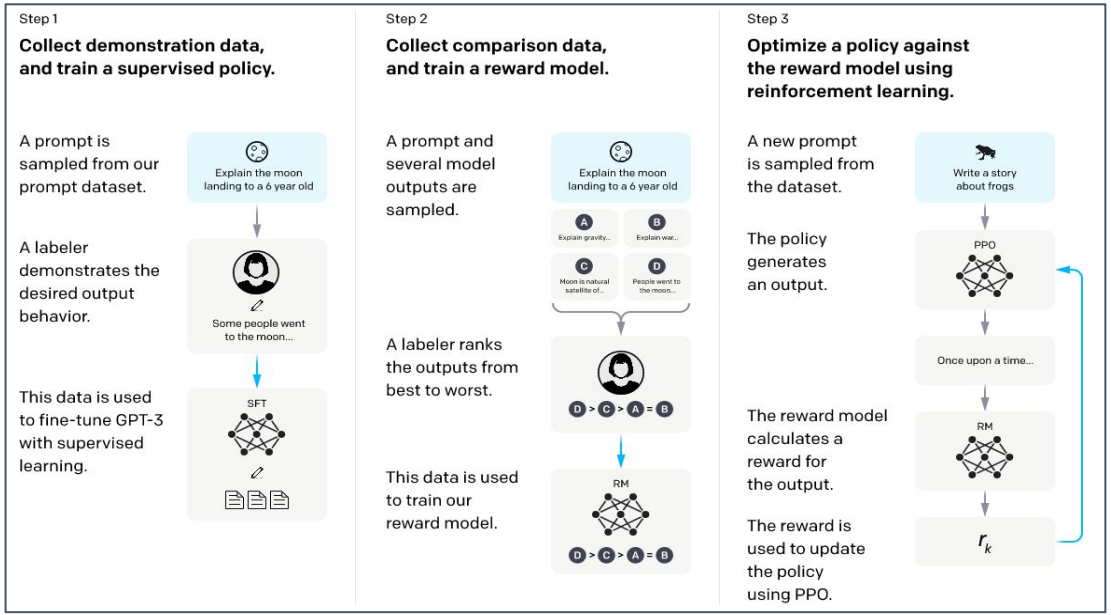
source: https://openai.com/research/instruction-following

InstructGPT & ChatGPT

Training language models to follow instructions with human feedback (Ouyang et al., 2022)

Nie powstała publikacja naukowa ani szczegółowy raport techniczny dotyczący **ChatGPT**, ale prawdopodobnie model był trenowany przy użyciu tych samych metod. Różnica wynika z użycia większej ilości danych dialogowych.

InstructGPT i ChatGPT nazywane są też **GPT-3.5**.



source: https://openai.com/research/instruction-following

GPT-4 and ChatGPT plugins

GPT-4 Technical Report (OpenAI, 2023)

14.03.2023 OpenAI ogłosiło publikację modelu **GPT-4:**

- opublikowano raport techniczny, ale większość informacji pozostaje nieujawniona
- by uzyskać dostęp należy opłacić subskrypcję
- na teraz udostępniona wersja GPT-4 wspiera tylko pracę z tekstem, ale wsparcie dla połączenia tekstu i obrazu (multimodal input) spodziewane jest w niedalekiej przyszłości
- Naukowcy z Microsoft ogłosili GPT-4 jako wczesną (i niepełną) wersję silnej sztucznej inteligencji (Artificial general intelligence - AGI)

GPT-4 and ChatGPT plugins

GPT-4 Technical Report (OpenAI, 2023)

14.03.2023 OpenAI ogłosiło publikację modelu **GPT-4:**

- opublikowano raport techniczny, ale większość informacji pozostaje nieujawniona
- by uzyskać dostęp należy opłacić subskrypcję
- na teraz udostępniona wersja GPT-4 wspiera tylko pracę z tekstem, ale wsparcie dla połączenia tekstu i obrazu (multimodal input) spodziewane jest w niedalekiej przyszłości
- Naukowcy z Microsoft ogłosili GPT-4 jako wczesną (i niepełną) wersję silnej sztucznej inteligencji (Artificial general intelligence - AGI)

W kolejnych dniach OpenAI ogłosił - **ChatGPT plugins**, zestaw narzędzi do ChatGPT pozwalający mu na dostęp do Internetu i interakcję z innymi narzędziami i serwisami, co ma pozwalać na dostęp do dodatkowej, aktualnej wiedzy i ma ograniczać problemy z halucynacjami.

Ogólnodostępne alternatywy

GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (Black et al., 2021)
GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model (Wang & Komatsuzaki, 2021)
GPT-NeoX-20B: An Open-Source Autoregressive Language Model (Black et al., 2022)
BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (Le Scao et al., 2022)
The Pile: An 800GB Dataset of Diverse Text for Language Modeling (Gao et al., 2021)

- **EleutherAI** - niezależni naukowcy starający się zreprodukować GPT-3 w jak największym stopniu:
  - **The Pile** - an 800GB zbiór danych testowych z treściami zebranymi z Internetu
  - **GPT-Neo** (2.7B) i **GPT-J** (6B) w 2021
  - **GPT-NeoX** (20B) w 2022
  - **Pythia** - zestaw modeli różnej wielkości, wytrenowany na tym samym zbiorze danych - 2023
- **togethercomputer** - platforma i zestaw modeli dostępnych do szerokiego użytku
  - **GPT-JT 6B** - GPT-J dotrenowany na naturalnych instrukcjach
  - **OpenChatKit** - modele bazujące na GPT-Neox and Pythia, dotrenowane na instrukcjach dialogowych
- **BigScience Workshop** - duża inicjatywa wielu naukowców z różnych instytucji
  - **BLOOM** (model wielojęzykowy, do 176B)

LLaMA i Alpaca

LLaMA: Open and Efficient Foundation Language Models (Touvron et al., 2023)
Alpaca: A Strong, Replicable Instruction-Following Model (Taori et al., 2023 - blogpost)

W końcówce lutego 2023, Meta opublikowało modele **LLaMA**:

- zbiór modeli od 7B do 65B parametrów
- wytrenowany na publicznie dostępnych danych
- licencja do użytku w projektach badawczych

Na początku marca 2023 r., naukowcy ze Stanford opublikowali model **Alpaca**:

- LLaMA 7B dotrenowany na 52k instrukcji wygenerowanych przez InstructGPT
- koszt - jedynie $600 (stworzenie zbioru danych: $500, 3 godziny treningu w chmurze: $100)
- w ślepym teście - podobna jakość do 175B text-davinci-003

Community-driven R&D

LoRA: Low-Rank Adaptation of Large Language Models (Hu et al., 2021)
LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale (Dettmers et al., 2022)

W ostatnich miesiącach kolejne inicjatywy i repozytoria z różnymi modelami językowymi powstają każdego dnia…

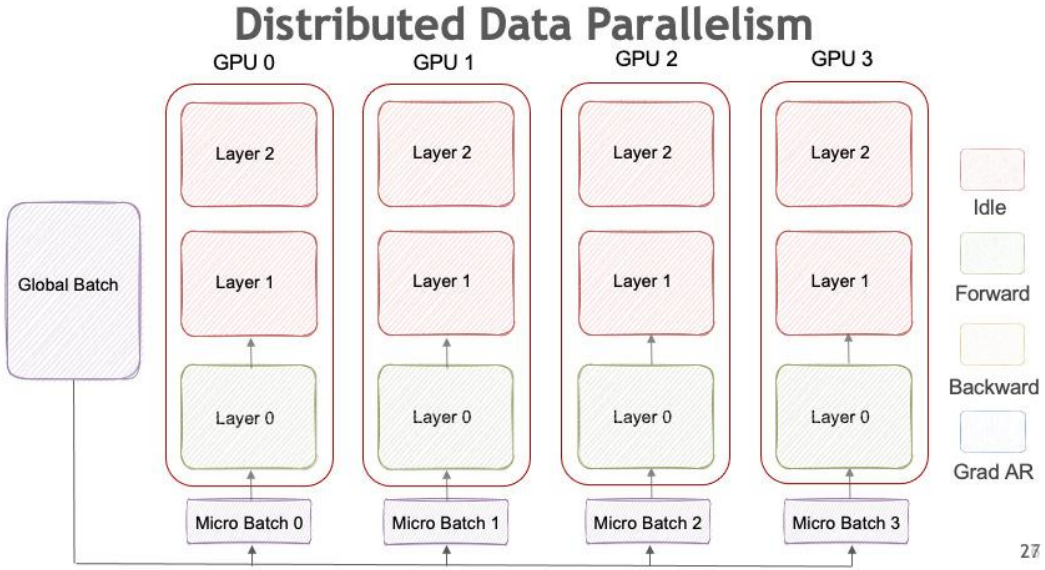
- alternatywy to ChatGPT: **Open Assistant**, **OpenChatKit**, **GPT4All**, **ChatLLaMA**, **Dolly, Dolly v2**
- zbiór syntetycznie wygenerowanych: **Open Instruction Generalist** (OIG; 43M instrukcji)
- uruchamianie (i trenowanie) modeli językowych na ograniczonym sprzęcie:
  - kwantyzacja: **LLaMA-int8**
  - porting modeli do C++: **LLaMA.cpp**
  - efektywny finetuning: **Alpaca-LoRA** (Low-Rank Adaptation)
- i wiele innych

deepsense.ai
BIG DATA SCIENCE

Metody optymalizacji wielkich modeli językowych

Data Parallelism

1. Zbiór danych dzielony jest na N części, gdzie N odpowiada liczbie dostępnych GPU. Każda z części przypisana jest do przydzielonego GPU.
2. Każde urządzenie przechowuje pełną kopię modelu i trenuje go na przydzielonym fragmencie danych.
3. Po backpropagacji gradienty zostają uśrednione i użyte do aktualizacji wag w każdej z kopii modelu.



Gif krok po kroku: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/_images/ddp.gif

Data Parallelism

Zalety

1. Łatwe w użyciu (np. domyślnie działa przy trenowaniu poprzez Trainera z HuggingFace).
2. Prędkość połączenia między urządzeniami nie jest tak istotna, jak w przypadku pozostałych metod zrównoleglania obliczeń.
3. Przyspiesza trening.

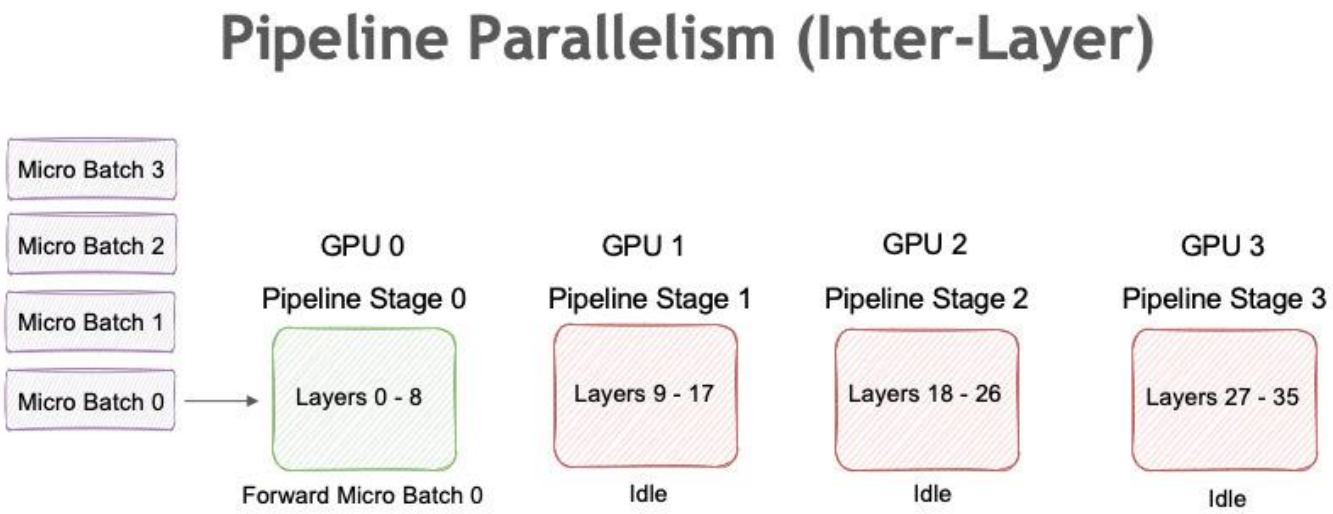
model	dataset	liczba GPU	liczba epok	czas treningu
flan-t5-base	SamSUM	1 GTX 1080 8GB	5	3h 53min
flan-t5-base	SamSUM	4 GTX 1080 8GB	5	**2h 15min** (1.7 razy szybciej)

Wady

1. Założenie, że model mieści się na pojedynczej GPU.

Pipeline Parallelism

Naiwne podejście

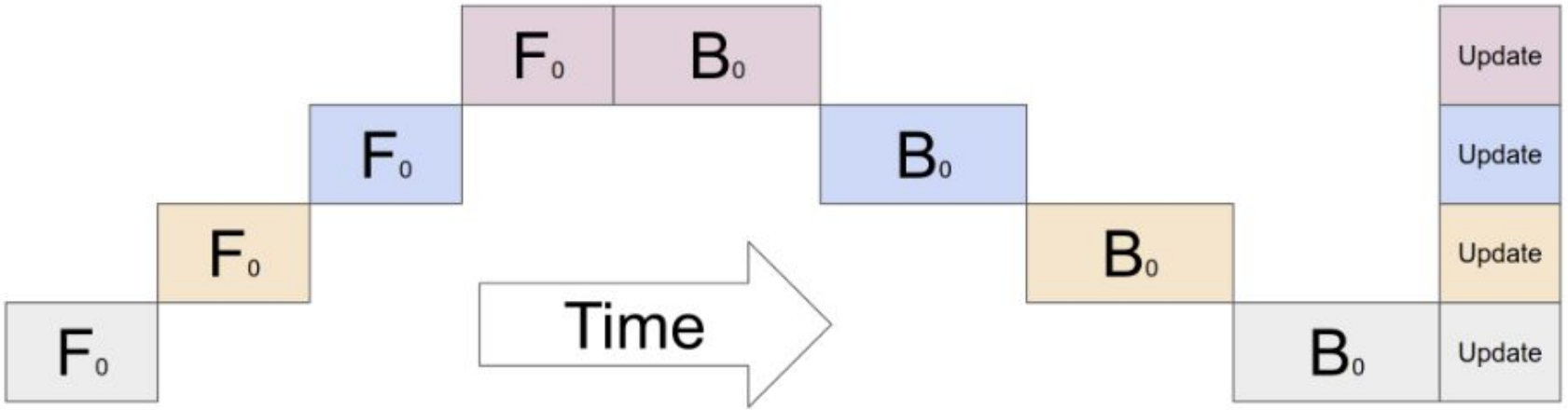


Źródło: https://docs.nvidia.com/deeplearning/nemo/user-guide/docs/en/stable/nlp/nemo_megatron/parallelisms.html

Pipeline Parallelism

Naiwne podejście

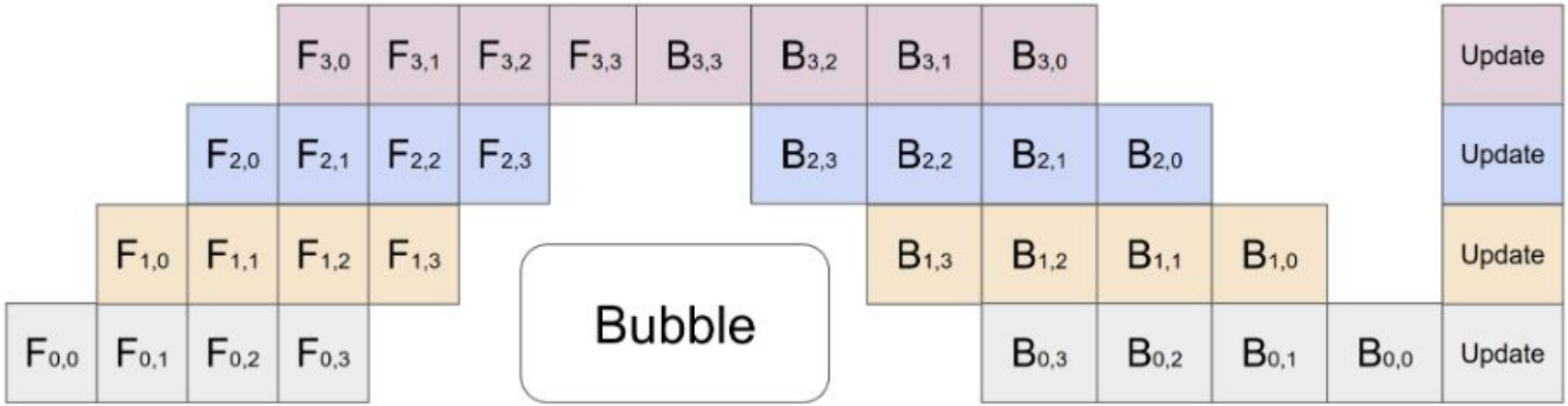
Pozwala na trenowanie modeli, które nie mieszczą się na pojedynczej karcie graficznej.
Problem z nieużywanymi kartami (idle time)



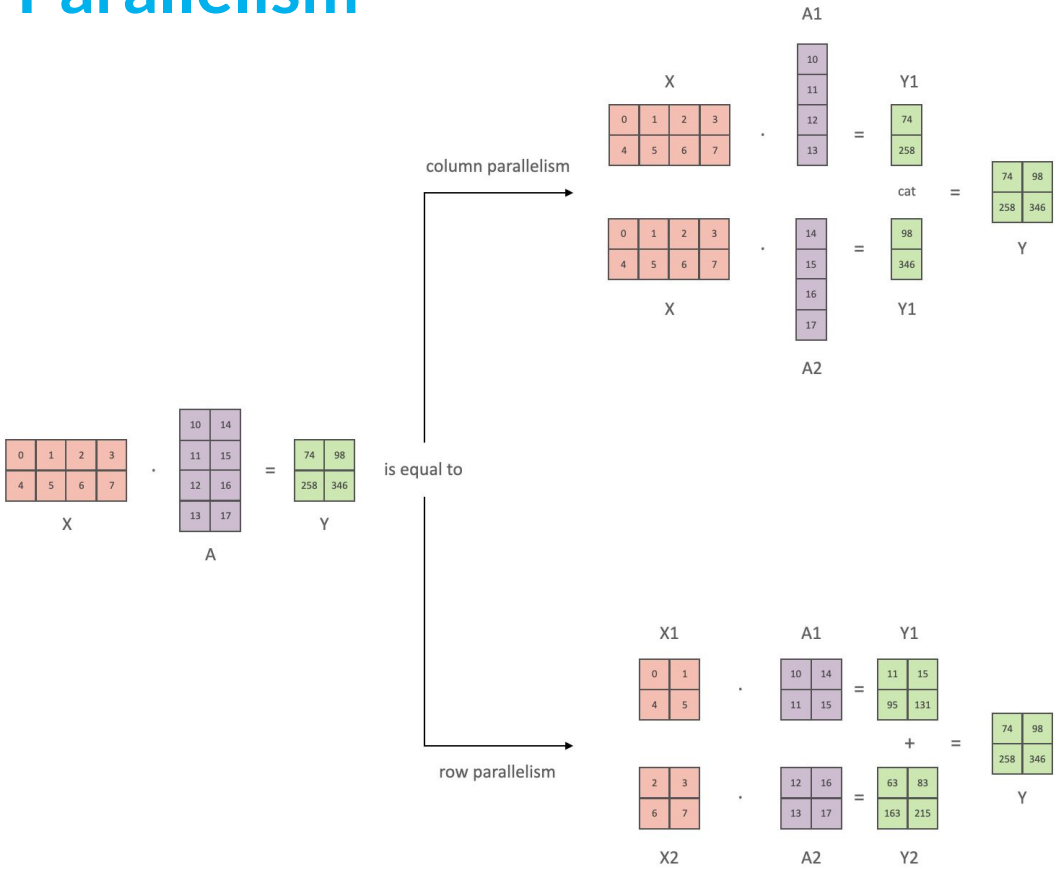
Pipeline Parallelism - micro batching

GPipe: Easy Scaling with Micro-Batch Pipeline Parallelism (Huang et al., 2019)

- Podział batcha na micro batche, aby w zrównoleglić obliczenia.
- Po propagacji wstecznej - średnia gradientów użyta jest do aktualizacji wag.
- Optymalna liczba micro batchy - 4N.

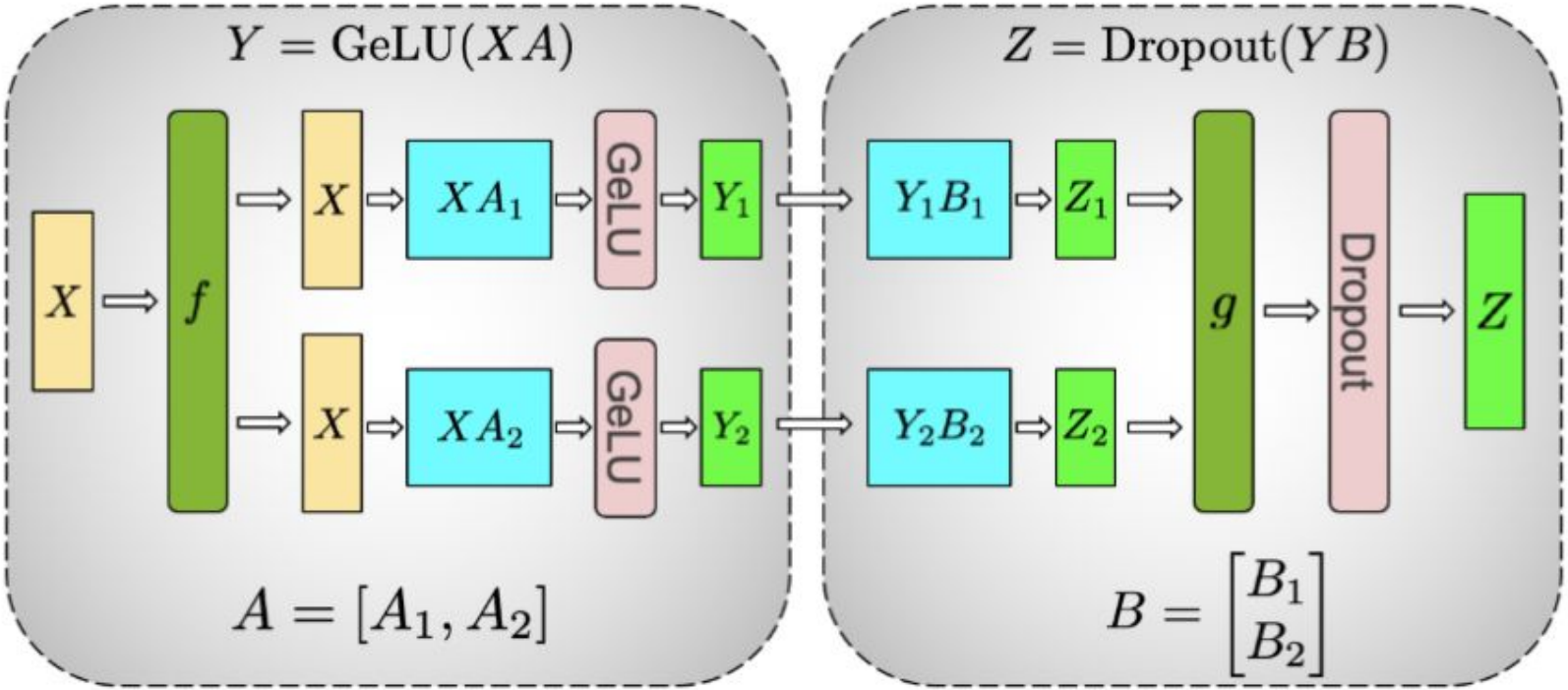


Tensor Parallelism



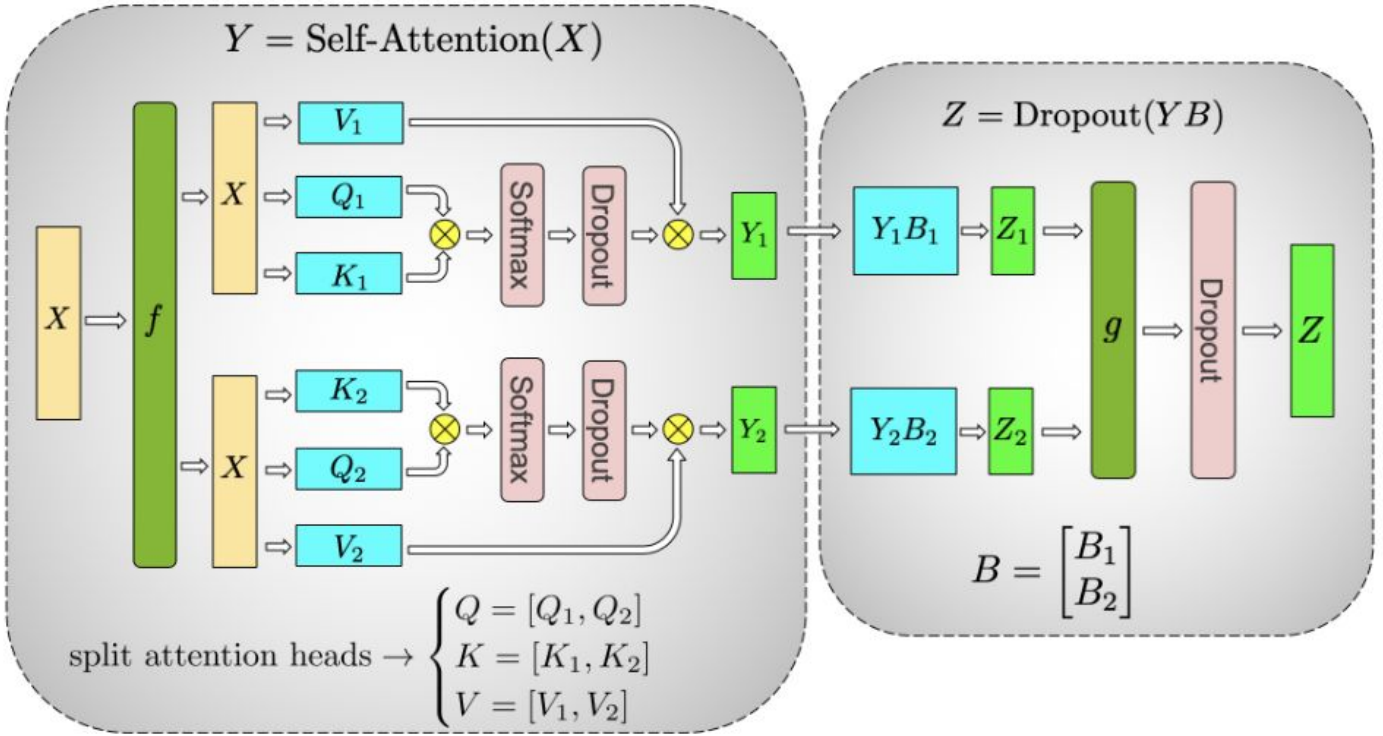
Tensor Parallelism - Multilayer Perceptron

Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism (Shoeybi et al., 2020)



Tensor Parallelism - Self-Attention

Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism (Shoeybi et al., 2020)

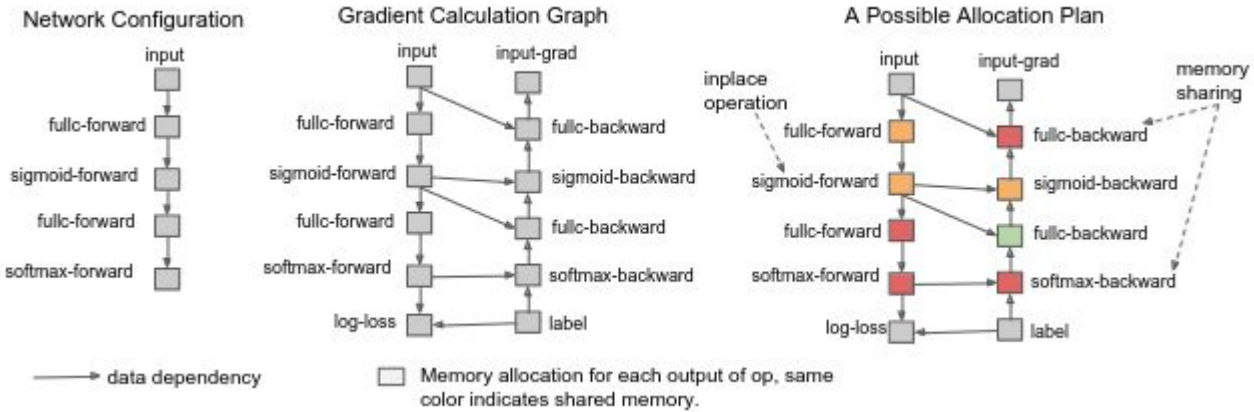


Activation Checkpointing

Training Deep Nets with Sublinear Memory Cost (Chen et al., 2016)

Metoda redukcji pamięci poprzez czyszczenie aktywacji poszczególnych warstw i liczenia ich ponownie podczas propagacji wstecznej.

Możemy ponownie alokować pamięć, która aktualnie jest niewykorzystywana co redukuje maksymalny peak zapotrzebowania pamięci przez sieć. Celem jest zmniejszenie zapotrzebowania na pamięć, ale kosztuje nas to czas potrzebny na ponowne obliczenia.



Mixed precision training

Sposoby reprezentacji liczb rzeczywistych

bfloat16: Brain Floating Point Format

Range: ~$1e^{-38}$ to ~$3e^{38}$

fp32: Single-precision IEEE Floating Point Format

Range: ~$1e^{-38}$ to ~$3e^{38}$

fp16: Half-precision IEEE Floating Point Format

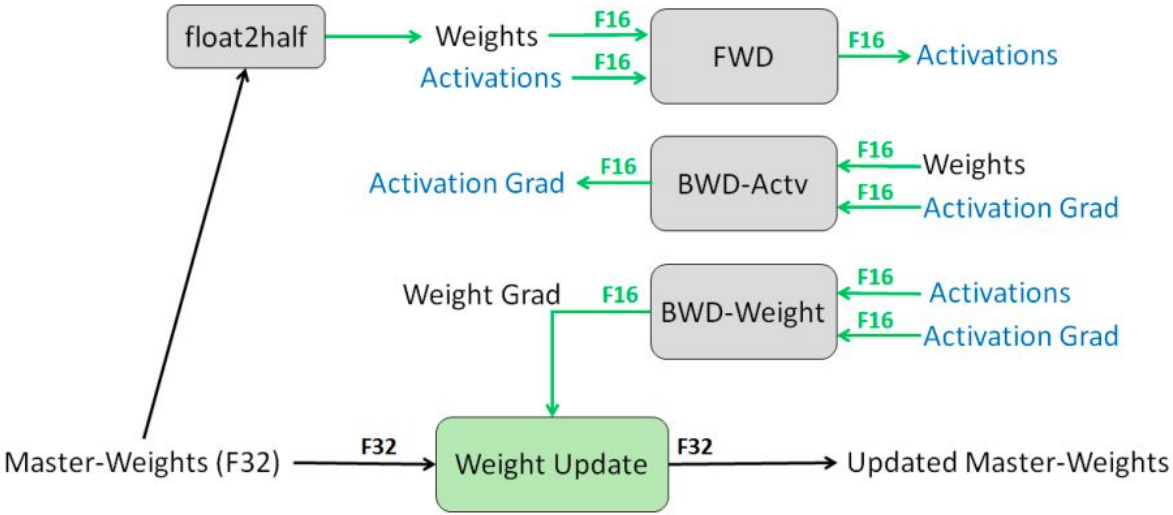
Range: ~$5.96e^{-8}$ to 65504

Mixed precision training

Mixed precision training (Narang et al., 2018)

W publikacji opisany jest sposób trenowania modeli w taki sposób, że wagi, aktywacje i gradienty przechowywane są jako FP16 (a nie FP32, jak wcześniej).

Żeby zachować jakość sieci z FP32 stosuje się kopię wag, która jest aktualizowana podczas kroku optimizera.



Mixed precision training

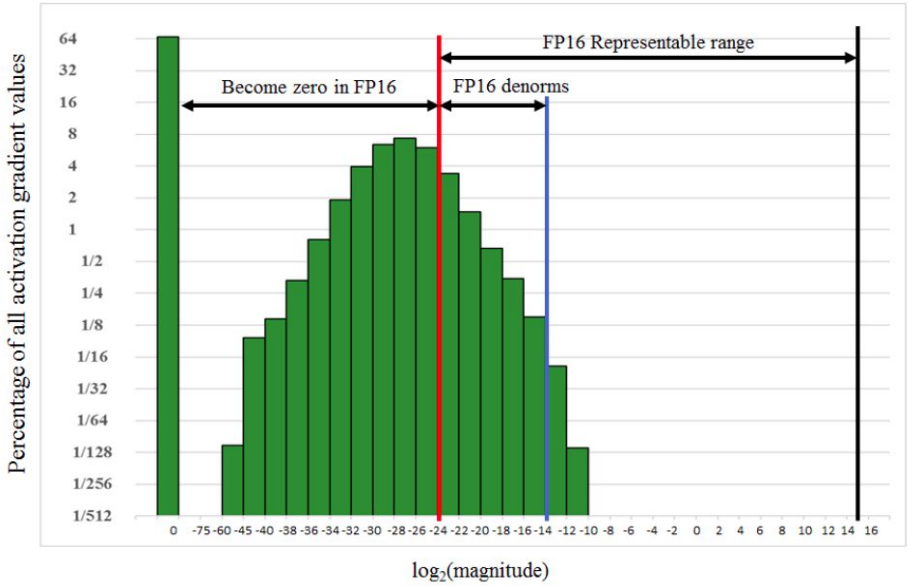
Loss Scaling

Przy zamianie gradientów na FP16 większość z nich zostanie przyrównana do 0.

Żeby temu zapobiec wprowadzone zostaje Loss Scaling.

Loss Scaling - loss jest przeskalowane przed propagacją wsteczną, wtedy dalej z reguły łańcucha gradienty są automatycznie przeskalowane.

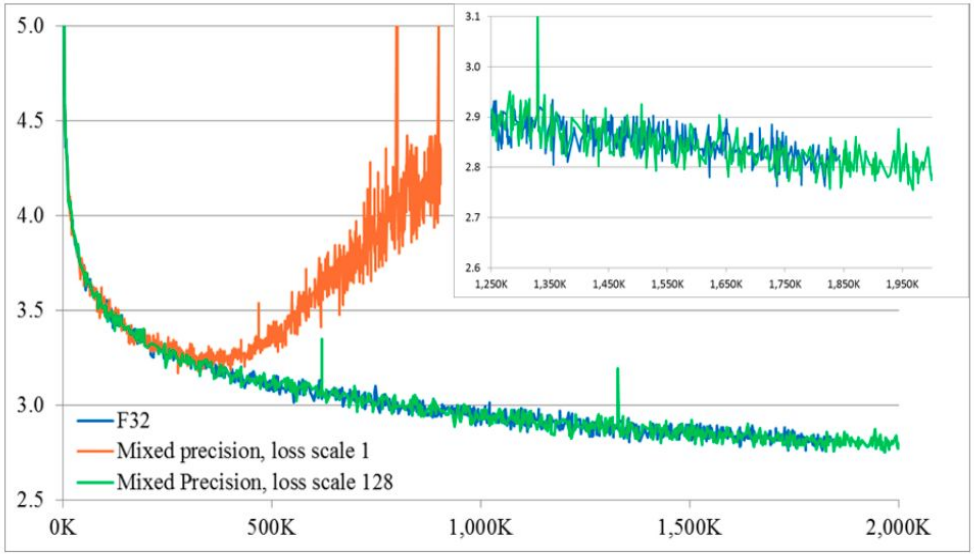
W publikacji raportują, że dla modelu Faster R-CNN dla detekcji obiektów osiągnęli w ten sposób accuracy takie same, jak używając FP32.



Mixed precision training

Wyniki dla modeli NLP

Model dla języka angielskiego, bigLSTM, trenowany na zbiorze złożonym z 1B słów o rozmiarze słownika 793k słów.



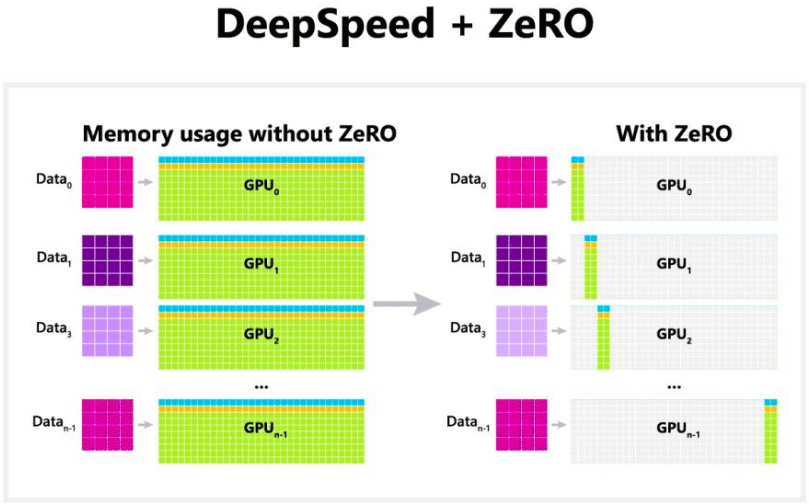
ZeRO

ZeRO: Memory Optimizations Toward Training Trillion Parameter Models (Rajbhandari et al., 2020)

DeepSpeed https://www.deepspeed.ai/

ZeRO (Zero Redundancy Optimizer)

Data Parallelism + metody zmniejszenia zużycia pamięci przez każde z GPU dzieląc różne stany modelu (wagi, gradienty i stany optymalizatora) na dostępne urządzenia (GPU i CPU).



ZeRO ma **trzy etapy optymalizacji**:

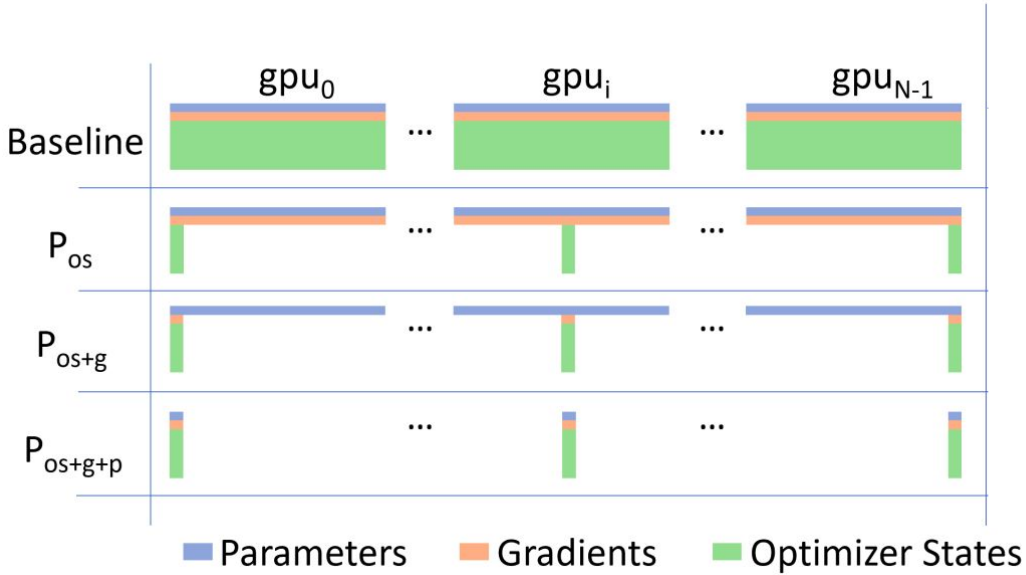
stage 1. - rozdzielanie stanów optimizera pomiędzy urządzenia

stage 2. - stage 1. + rozdzielanie gradientów

stage 3. - stage 1. + stage 2. + rozdzielanie parametrów modelu

ZeRO

Ilustracja działania



ZeRO

Wyniki

DP	**7.5B Model** (GB)			**128B Model** (GB)			**1T Model** (GB)		
	P_{os}	P_{os+g}	P_{os+g+p}	P_{os}	P_{os+g}	P_{os+g+p}	P_{os}	P_{os+g}	P_{os+g+p}
1	120	120	120	2048	2048	2048	16000	16000	16000
4	52.5	41.3	**30**	896	704	512	7000	5500	4000
16	35.6	**21.6**	7.5	608	368	128	4750	2875	1000
64	**31.4**	16.6	1.88	536	284	**32**	4187	2218	250
256	30.4	15.4	0.47	518	263	8	4046	2054	62.5
1024	30.1	15.1	0.12	513	257	2	4011	2013	**15.6**

Table 1: Per-device memory consumption of different optimizations in *ZeRO*-DP as a function of DP degree . Bold-faced text are the combinations for which the model can fit into a cluster of 32GB V100 GPUs.

Dostępne narzędzia

- **DeepSpeed** - https://github.com/microsoft/DeepSpeed
  - implementacja ZeRO (może być użyte bezpośrednio w Trainerze HuggingFace - wystarczy odpowiedni config)
  - wspiera trening z Data/Pipeline Parallelism, MegatronLM i mixed precision
- **accelerate** - https://github.com/huggingface/accelerate
  - mixed precision
  - ZeRO (z własną pętlą treningową, nie bezpośrednio przez Trainera)
  - zrównoleglona inferencja
  - MegatronLM (DDP, PP, TP, mixed precision, głównie dla GPT-2)
- **MegatronLM** - https://github.com/NVIDIA/Megatron-LM
  - działa dla architektur T5, GPT-2 i BERT
  - TP, PP, MP
  - checkpointy są zapisywane w rozproszonym formacie

Optymalizacja modeli przy wdrożeniu

Przyspieszenie inferencji

Quantization	• Zmniejszenie precyzji parametrów modelu np. f32 → f16 • Może znacząco wpływać na jakość modelu, dlatego stosuje się Quantization Aware Training (QAT)
Pruning	• Unstructured & structured pruning • Kryterium pruningu
Knowledge Distillation	• Transfer wiedzy z dużego modelu do mniejszego • Rodzaje knowledge distillation
Optimizing for hardware	• TesorRT

deepsense.ai
BIG DATA SCIENCE

Dziękujemy!

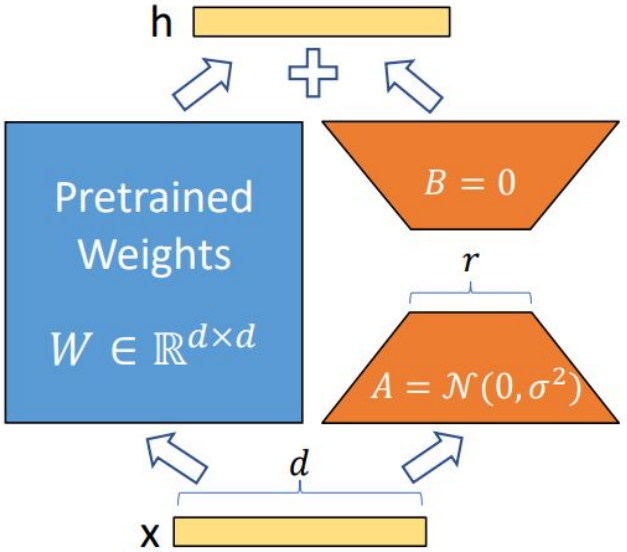
Adam Maciaszek
Szymon Janowski
credits to
Artur Zygadło
Alicja Kotyla

LoRA

LoRA: Low-Rank Adaptation of Large Language Models(Hu et al., 2021)

Metoda efektywnego finetuningu modeli

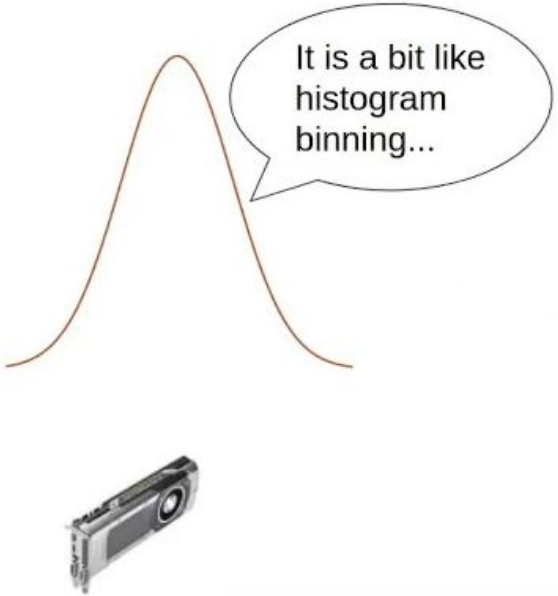
- Zamrożenie wag modelu
- Trening ograniczonej liczby parametrów i nałożenie dotrenowanych wag na podstawowy model
- Brak konieczności przechowywania gradientów i stanów optymalizatora
- Brak konieczności zapisu całego modelu
- Możliwość łatwego przełączania się pomiędzy różnymi wersjami modelu

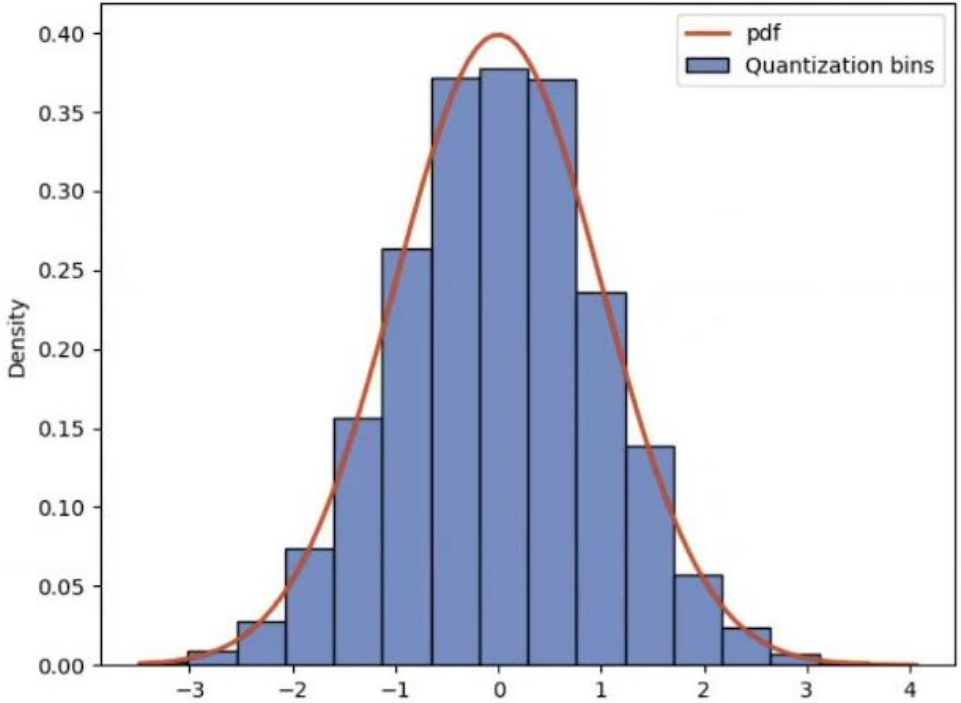


8-bit optimizers

8-bit Optimizers via Block-wise Quantization (Dettmers et al., 2022)

How does quantization work?





Ciekawe materiały

1. BigScience BLOOM | 3D Parallelism Explained | Large Language Models | ML Coding Series, Aleksa Gordić - The AI Epiphany
   https://youtu.be/pTChDs5uD8I
2. Ultimate Guide To Scaling ML Models - Megatron-LM | ZeRO | DeepSpeed | Mixed Precision, Aleksa Gordić - The AI Epiphany
   https://youtu.be/hc0u4avAkuM
3. Blog post o DeepSpeed i ZeRO
   https://www.microsoft.com/en-us/research/blog/zero-deepspeed-new-system-optimizations-enable-training-models-with-over-100-billion-parameters/
4. The Technology Behind BLOOM Training
   https://huggingface.co/blog/bloom-megatron-deepspeed
5. Publikacja nt. ZeRO
   https://arxiv.org/pdf/1910.02054.pdf
6. Publikacja nt. Mixed Precision Training
   https://arxiv.org/pdf/1710.03740.pdf
7. Blog post o formacie BF16
   https://cloud.google.com/blog/products/ai-machine-learning/bfloat16-the-secret-to-high-performance-on-cloud-tpus