SEZNAM.CZ

Velké jazykové modely

Učení a provoz

Jan Petrov

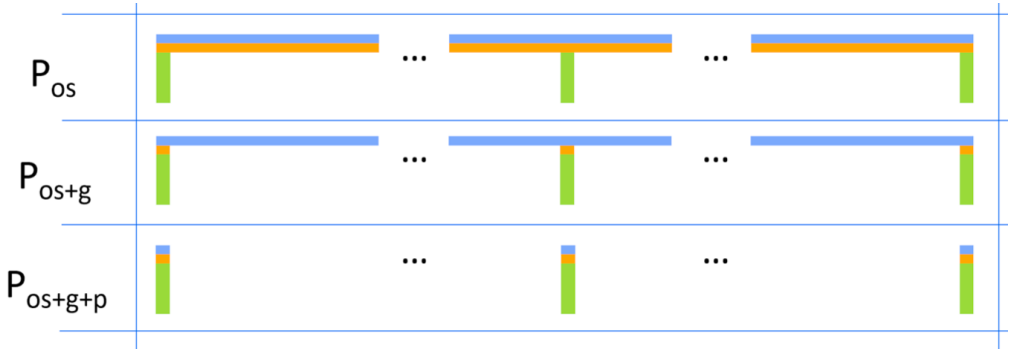
Senior Research Engineer, LLM tým

Když jedno GPU nestačí

Jeden node, více GPUs

- Až 18B na parametr modelu + aktivace
- Gradienty → Optimizér → Parametry modelu
- 7B → 126 GB + aktivace | 8 GPUs: 8 × 126 GB?

DeepSpeed Zero



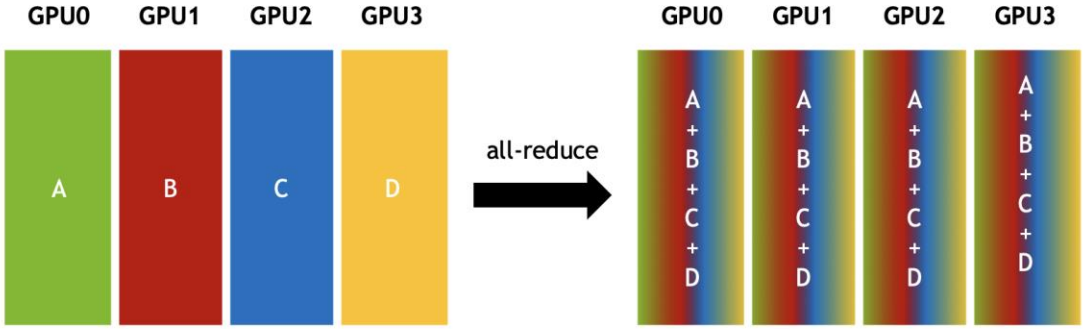
https://arxiv.org/pdf/1910.02054

Více nodů

- 100 – 400 gbit | ne TCP/IP packety | RoCE, Mellanox

Synchronizace

- Nvidia Collective Communications Library (NCCL)
- all_reduce a další základní operace

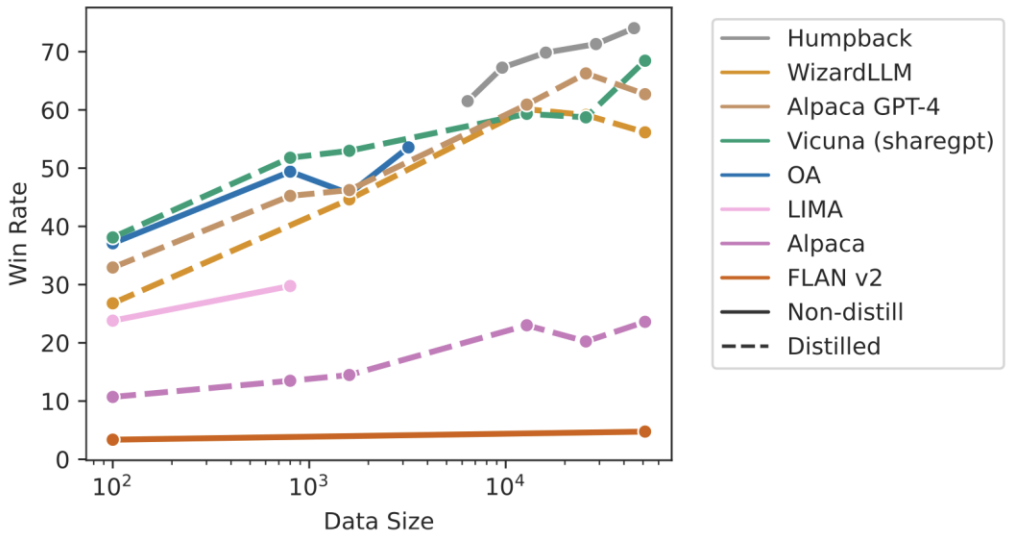


https://images.nvidia.com/events/sc15/pdfs/NCCL-Woolley.pdf

Dodatečné učení

Instrukční finetuning

- SFT = supervised finetuning
- Vzorová otázka – odpověď, vzorová konverzace
- Důraz na kvalitu, logaritmický vliv kvantity



Self-Alignment with Instruction Backtranslation, ICLR 2024

Rejection sampling

- Různé odpovědi na 1 prompt
- Výběr nejlepší a nejhorší

Reward model

- Model hodnotící kvalitu výstupu
- Často učený na pairwise datech
- Nutná komponenta RLHF

DPO

- (Prompt, lepší odpověď, horší odpověď)
- Reward model | lidská anotace | LLM anotace
- Rejection sampling | vylepšení odpovědi

Syntetická data

Význam

- Strmý růst popularity, mnoho postupů
- Přimíchávání ve fázi předtrénování (pretraining)
- NVidia Nemotron 340B, Llama 3.1, …

Uplatnění

- Matematika, kód | dlouhý kontext | instrukce | chat | function calling | alignment …

Přístupy

Prosté dotazování

- Přimíchávání ve fázi předtrénování (pretraining)

Iterativní zvyšování složitosti

- WizardLM

Transformace

- Dotaz „chceme xml“ a výstup do xml

Reverzní přístupy

- Máme text a generujeme k němu otázky

Uvažování ve více krocích, agenti, …

Hodnocení kvality

Lidské evaluace

- Výzkum
- Produktoví manažeři pro konkrétní použití na cílové úlohy, např. sumarizace SERP

LLM evaluace

- Absolutní vs párové anotace
- Kategorie kvality, čeština
- Prompty se vzorovými odpověďmi
- Korelace s lidskými anotacemi

Automatické

Szn: cílová délka textu, extrakce pojmu, …
Dlouhý kontext: needle in haystack apod.

Harness: Přeložené a vlastní benchmarky

novinka: CZLC/BenCzechMark

Následují otázky s výběrem jediné správné odpovědi týkající se středoškolské statistiky.
{ukázky}
Který z následujících předpokladů se neuplatňuje v případě binomického rozdělení?
A. Každý pokus je považován buď za úspěch, nebo za neúspěch.
B. Každý pokus je nezávislý.
C. Hodnota zkoumané náhodné proměnné je počet pokusů do prvního úspěchu.
D. Je stanoven pevný počet pokusů.
Odpověď:

Speciální tokeny, formáty a funkcionality

EOS token

```
<|begin_of_text|>. .
.<|end_of_text|>
```

Chat

```
<|begin_of_text|>
<|start_header>user<|end_header|>
vstup od uživatele
<|start_header>assistant<end_header|>
```

JSON, XML

```
{"kvalita": 3, "jazyk": "cs",
"alignment": true}
```

Function calling

```
<|begin_of_text|>
<|start_header>system<|end_header|>
Jsi užitečný asistent s přístupem k
následujícím funkcím: [{name: "calculate",
"parameters":..., "description":...},...]
<start_header>user<end_header|>
Ahoj, mám 27 palet po 33 kusech, kolik mám
kusů celkem?
<start_header>assistant<end_header|>
```

Speciální tokeny, formáty a funkcionality

EOS token

```
<|begin_of_text|>. .
.<|end_of_text|>
```

Chat

```
<|begin_of_text|>
<|start_header>user<|end_header|>
vstup od uživatele
<|start_header>assistant<end_header|>
```

JSON, XML

```
{"kvalita": 3, "jazyk": "cs",
"alignment": true}
```

Function calling

```
<|begin_of_text|>
<|start_header>system<|end_header|>
Jsi užitečný asistent s přístupem k
následujícím funkcím: [{name: "calculate",
"parameters":..., "description":...},...]
<start_header>user<end_header|>
Ahoj, mám 27 palet po 33 kusech, kolik mám
kusů celkem?
<start_header>assistant<end_header|>
<start_header>call<end_header|>
{name: "calculate", text_input: "27*33"}
<|eot|>
```

Speciální tokeny, formáty a funkcionality

EOS token

```
<|begin_of_text|>. .
.<|end_of_text|>
```

Chat

```
<|begin_of_text|>
<|start_header>user<|end_header|>
vstup od uživatele
<|start_header>assistant<end_header|>
```

JSON, XML

```
{"kvalita": 3, "jazyk": "cs",
"alignment": true}
```

Function calling

```
<|begin_of_text|>
<|start_header>system<|end_header|>
Jsi užitečný asistent s přístupem k
následujícím funkcím: [{name: "calculate",
"parameters":..., "description":...},...]
<start_header>user<end_header|>
Ahoj, mám 27 palet po 33 kusech, kolik mám
kusů celkem?
<start_header>assistant<end_header|>
<start_header>call<end_header|>
{name: "calculate", text_input: "27*33"}
<|eot|>
<start_header>response<end_header|>
{status: True, result: 891}
```

Speciální tokeny, formáty a funkcionality

EOS token

```
<|begin_of_text|>. .
.<|end_of_text|>
```

Chat

```
<|begin_of_text|>
<|start_header>user<|end_header|>
vstup od uživatele
<|start_header>assistant<end_header|>
```

JSON, XML

```
{"kvalita": 3, "jazyk": "cs",
"alignment": true}
```

Function calling

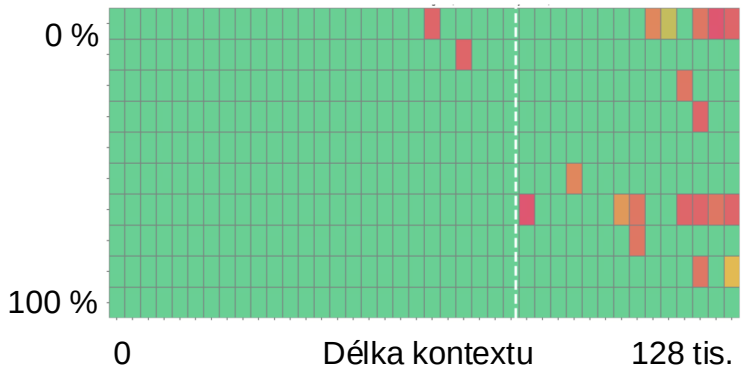
```
<|begin_of_text|>
<|start_header>system<|end_header|>
Jsi užitečný asistent s přístupem k
následujícím funkcím: [{name: "calculate",
"parameters":..., "description":...},...]
<start_header>user<end_header|>
Ahoj, mám 27 palet po 33 kusech, kolik mám
kusů celkem?
<start_header>assistant<end_header|>
<start_header>call<end_header|>
{name: "calculate", text_input: "27*33"}
<|eot|>
<start_header>response<end_header|>
{status: True, result: 891}
<start_header>assistant<end_header|>
Máte celkem 891 kusů.<|eot|>
```

Dlouhý kontext

Llama 2	**4096**
Llama 3	**8192**
Llama 3.1	**128 000**
GPT 4o	**128 000**
Claude	**200 000**
Gemini	**2M**

Automatické měření kvality

Needle in haystack a další testy



Kvalita sumarizací

Povrchně hezké, ale možná nepravdivé

FABLES: claude 90.66, gpt4o 78.16, mixtral 70.4

Dlouhý kontext

Velké nároky

- Pouhé 4 vzorky plné délky mohou zabrat více GPU paměti než parametry modelu
- Výpočetní náročnost

Možnosti zmenšení

- Kvantizace 2B → 1B
- Grouped query attention (Llama 3 i pro 8B model)
- Sliding window (Mistral 0.1, moc nefunguje)
- Aktivní výzkum (zatím ne do produkce)

KV cache (Llama 70)

vrstvy	**80**
dimenze	**8192**
fp16 (2 bajty)	**2**
key + value	**2**
tokenů	**128 000**
grouped query attention	***/ 8***
CELKEM	**42 GB**

Inference výstupů

Inferenční prostředí

- Naivní Transformers inference
- vLLM (různí výrobci) a další
- TensorRT-LLM (CUDA, NVidia), Triton server

Serving

- Throughput vs latence (TTFT = time to first token)
- Inflight batching
- 70 GB llama fp16 – 4× 80 GB GPU

Kvantizace

- fp16 (bf16) → fp8 (H100)
- int 4, 1 bit | produkční efektivita vs. kapacita karty
- Vliv na kvalitu
- Post-training kalibrace, trénování zohledňující kvantizaci

Nákladová efektivnost

- Méně karet + rychlejší fp8 výpočet → např. 3× levnější
- Různá nastavení kvantizace u komerčních služeb
- Škálování cloudu
- Peak vs. dávkové úlohy

Jan Petrov

Senior Research Engineer, LLM tým

jan.petrov@firma.seznam.cz