unsloth/DeepSeek-R1-Distill-Qwen-14B-unsloth-bnb-4bit

like 17 Follow Unsloth AI 2.51k

Text Generation · Transformers · Safetensors · English · qwen2 · deepseek · qwen · unsloth · conversational · text-generation-inference · Inference Endpoints · 4-bit precision · bitsandbytes · License: apache-2.0

Train · Deploy · Use this model

Model card · Files · Community

See our collection for versions of Deepseek-R1 including GGUF and original formats.

Dynamic 4-bit: Unsloth's Dynamic 4-bit Quants selectively avoids quantizing certain parameters, greatly increase accuracy than standard 4-bit. See our full collection of Unsloth quants on Hugging Face here.

Finetune LLMs 2-5x faster with 70% less memory via Unsloth!

We have a free Google Colab Tesla T4 notebook for Llama 3.1 (8B) here: https://colab.research.google.com/github/unslothai/notebooks/blob/main/nb/Llama3.1_(8B)-Alpaca.ipynb

Join our Discord

made with unsloth

Downloads last month
16,138

Safetensors

Model size	9.78B params
Tensor type	F32 · BF16 · U8

Inference Providers NEW

Text Generation

This model is not currently available via any of the supported third-party Inference Providers, and the model is not deployed on the HF Inference API.

Model tree for unsloth/DeepSeek-...

Base model	deepseek-ai/DeepSeek-...
Quantized (87)	this model
Finetunes	12 models
Quantizations	5 models

Spaces using unsloth/DeepSeek... 3

KBaba7/Quant · totolook/Quant · ruslanmv/convert_to_gguf

✨ Finetune for Free

All notebooks are **beginner friendly**! Add your dataset, click "Run All", and you'll get a 2x faster finetuned model which can be exported to GGUF, vLLM or uploaded to Hugging Face.

Unsloth supports	Free Notebooks	Performance	Memory use
Llama-3.2 (3B)	Start on Colab	2.4x faster	58% less
Llama-3.2 (11B vision)	Start on Colab	2x faster	60% less
Qwen2 VL (7B)	Start on Colab	1.8x faster	60% less
Qwen2.5 (7B)	Start on Colab	2x faster	60% less
Llama-3.1 (8B)	Start on Colab	2.4x faster	58% less
Phi-3.5 (mini)	Start on Colab	2x faster	50% less
Gemma 2 (9B)	Start on Colab	2.4x faster	58% less
Mistral (7B)	Start on Colab	2.2x faster	62% less

Documentation

- This Llama 3.2 conversational notebook is useful for ShareGPT ChatML / Vicuna templates.
- This text completion notebook is for raw text. This DPO notebook replicates Zephyr.

Collections including unsloth/Deep…

Unsloth 4-bit Dynamic … Collection
Unsloths Dyna… • 20 items • U. • △ 41

DeepSeek R1 (All Versi… Collection
DeepSeek R1 -… • 29 items • U. • △ 163

- * Kaggle has 2x T4s, but we use 1. Due to overhead, 1x T4 is 5x faster.

Special Thanks

A huge thank you to the DeepSeek team for creating and releasing these models.

1. Introduction

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrated remarkable performance on reasoning. With RL, DeepSeek-R1-Zero naturally emerged with numerous powerful and interesting reasoning behaviors. However, DeepSeek-R1-Zero encounters challenges such as endless repetition, poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1 across math, code, and reasoning tasks. To support the research community, we have open-sourced DeepSeek-R1-Zero, DeepSeek-R1, and six dense models distilled from DeepSeek-R1 based on Llama and Qwen. DeepSeek-R1-Distill-Qwen-32B outperforms OpenAI-o1-mini across various benchmarks, achieving new state-of-the-art results for dense models.

2. Model Summary

Post-Training: Large-Scale Reinforcement Learning on the Base Model

- We directly apply reinforcement learning (RL) to the base model without relying on supervised fine-tuning (SFT) as a preliminary step. This approach allows the model to explore chain-of-thought (CoT) for solving complex problems, resulting in the development of DeepSeek-R1-Zero. DeepSeek-R1-Zero demonstrates capabilities such as self-verification, reflection, and generating long CoTs, marking a significant milestone for the research community. Notably, it is the first open research to validate that reasoning capabilities of LLMs can be incentivized purely through RL, without the need for SFT. This breakthrough paves the way for future advancements in this area.

- We introduce our pipeline to develop DeepSeek-R1. The pipeline incorporates two RL stages aimed at discovering improved reasoning patterns and aligning with human preferences, as well as two SFT stages that serve as the seed for the model's reasoning and non-reasoning capabilities. We believe the pipeline will benefit the industry by creating better models.

Distillation: Smaller Models Can Be Powerful Too

- We demonstrate that the reasoning patterns of larger models can be distilled into smaller models, resulting in better performance compared to the reasoning patterns discovered through RL on small models. The open source DeepSeek-R1, as well as its API, will benefit the research community to distill better smaller models in the future.

- Using the reasoning data generated by DeepSeek-R1, we fine-tuned several dense models that are widely used in the research community. The evaluation results demonstrate that the distilled smaller dense models perform exceptionally well on benchmarks. We open-source distilled 1.5B, 7B, 8B, 14B, 32B, and 70B checkpoints based on Qwen2.5 and Llama3 series to the community.

3. Model Downloads

DeepSeek-R1 Models

Model	#Total Params	#Activated Params	Context Length	Downloa
DeepSeek-R1-Zero	671B	37B	128K	🤗 HuggingFa
DeepSeek-R1	671B	37B	128K	🤗 HuggingFa

DeepSeek-R1-Zero & DeepSeek-R1 are trained based on DeepSeek-V3-Base. For more details regrading the model architecture, please refer to DeepSeek-V3 repository.

DeepSeek-R1-Distill Models

Model	Base Model	Download
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-Math-1.5B	🤗 HuggingFace

Model	Base Model	Download
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B	🤗 HuggingFace
DeepSeek-R1-Distill-Llama-8B	Llama-3.1-8B	🤗 HuggingFace
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B	🤗 HuggingFace
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B	🤗 HuggingFace
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-Instruct	🤗 HuggingFace

DeepSeek-R1-Distill models are fine-tuned based on open-source models, using samples generated by DeepSeek-R1. We slightly change their configs and tokenizers. Please use our setting to run these models.

4. Evaluation Results

DeepSeek-R1-Evaluation

For all our models, the maximum generation length is set to 32,768 tokens. For benchmarks requiring sampling, we use a temperature of 0.6, a top-p value of 0.95, and generate 64 responses per query to estimate pass@1.

Category	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	Architecture	-	-	MoE

Category	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	# Activated Params	-	-	37B
	# Total Params	-	-	671B
English	MMLU (Pass@1)	88.3	87.2	88.5
	MMLU-Redux (EM)	88.9	88.0	89.1
	MMLU-Pro (EM)	78.0	72.6	75.9
	DROP (3-shot F1)	88.3	83.7	91.6
	IF-Eval (Prompt Strict)	**86.5**	84.3	86.1
	GPQA-Diamond (Pass@1)	65.0	49.9	59.1
	SimpleQA (Correct)	28.4	38.2	24.9
	FRAMES (Acc.)	72.5	80.5	73.3
	AlpacaEval2.0 (LC-winrate)	52.0	51.1	70.0
	ArenaHard (GPT-4-1106)	85.2	80.4	85.5
Code	LiveCodeBench (Pass@1-COT)	33.8	34.2	-
	Codeforces (Percentile)	20.3	23.6	58.7
	Codeforces (Rating)	717	759	1134

Category	Benchmark (Metric)	Claude-3.5-Sonnet-1022	GPT-4o 0513	DeepSeek V3
	SWE Verified (Resolved)	**50.8**	38.8	42.0
	Aider-Polyglot (Acc.)	45.3	16.0	49.6
Math	AIME 2024 (Pass@1)	16.0	9.3	39.2
	MATH-500 (Pass@1)	78.3	74.6	90.2
	CNMO 2024 (Pass@1)	13.1	10.8	43.2
Chinese	CLUEWSC (EM)	85.4	87.9	90.9
	C-Eval (EM)	76.7	76.0	86.5
	C-SimpleQA (Correct)	55.4	58.7	**68.0**

Distilled Model Evaluation

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1
GPT-4o-0513	9.3	13.4	74.6	49.9
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0
o1-mini	63.6	80.0	90.0	60.0

Model	AIME 2024 pass@1	AIME 2024 cons@64	MATH-500 pass@1	GPQA Diamond pass@1
QwQ-32B-Preview	44.0	60.0	90.6	54.5
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**

5. Chat Website & API Platform

You can chat with DeepSeek-R1 on DeepSeek's official website: chat.deepseek.com, and switch on the button "DeepThink"

We also provide OpenAI-Compatible API at DeepSeek Platform: platform.deepseek.com

6. How to Run Locally

DeepSeek-R1 Models

Please visit DeepSeek-V3 repo for more information about running DeepSeek-R1 locally.

DeepSeek-R1-Distill Models

DeepSeek-R1-Distill models can be utilized in the same manner as Qwen or Llama models.

For instance, you can easily start a service using vLLM:

```
vllm serve deepseek-ai/DeepSeek-R1-Distill
```

NOTE: We recommend setting an appropriate temperature (between 0.5 and 0.7) when running these models, otherwise you may encounter issues with endless repetition or incoherent output.

7. License

This code repository and the model weights are licensed under the MIT License. DeepSeek-R1 series support commercial use, allow for any modifications and derivative works, including, but not limited to, distillation for training other LLMs. Please note that:

- DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B and DeepSeek-R1-Distill-Qwen-32B are derived from Qwen-2.5 series, which are originally licensed under Apache 2.0 License, and now finetuned with 800k samples curated with DeepSeek-R1.

- DeepSeek-R1-Distill-Llama-8B is derived from Llama3.1-8B-Base and is originally licensed under llama3.1 license.
- DeepSeek-R1-Distill-Llama-70B is derived from Llama3.3-70B-Instruct and is originally licensed under llama3.3 license.

8. Citation

9. Contact

If you have any questions, please raise an issue or contact us at service@deepseek.com.