Skip to main content

[![Cornell University](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

In just 3 minutes help us improve arXiv:

[Annual Global Survey](https://cornell.ca1.qualtrics.com/jfe/form/SV_6m22mbqW9GQ3pQO)

We gratefully acknowledge support from the Simons Foundation, [member institutions](https://info.arxiv.org/about/ourmembers.html), and all contributors. [Donate](https://info.arxiv.org/about/donate.html)

[](/IgnoreMe)

[![arxiv logo](/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg)](/) > [cs](/list/cs/recent) > arXiv:2210.17323

[![arXiv logo](/static/browse/0.3.4/images/arxiv-logomark-small-white.svg)](https://arxiv.org/)

[![Cornell University Logo](/static/browse/0.3.4/images/icons/cu/cornell-reduced-white-SMALL.svg)](https://www.cornell.edu/)

open search

GO

open navigation menu

quick links

* [Login](https://arxiv.org/login)
* [Help Pages](https://info.arxiv.org/help)
* [About](https://info.arxiv.org/about)

Computer Science > Machine Learning

arXiv:2210.17323 (cs)

[Submitted on 31 Oct 2022 ([v1](https://arxiv.org/abs/2210.17323v1)), last revised 22 Mar 2023 (this version, v2)]

Title:GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers

Authors:[Elias

Frantar](https://arxiv.org/search/cs?searchtype=author&query=Frantar,+E),

[Saleh

Ashkboos](https://arxiv.org/search/cs?searchtype=author&query=Ashkboos,+S),

[Torsten

Hoefler](https://arxiv.org/search/cs?searchtype=author&query=Hoefler,+T), [Dan

Alistarh](https://arxiv.org/search/cs?searchtype=author&query=Alistarh,+D)

View a PDF of the paper titled GPTQ: Accurate Post-Training Quantization for

Generative Pre-trained Transformers, by Elias Frantar and 3 other authors

[View PDF](/pdf/2210.17323)

> Abstract:Generative Pre-trained Transformer models, known as GPT or OPT, set
> themselves apart through breakthrough performance across complex language
> modelling tasks, but also by their extremely high computational and storage
> costs. Specifically, due to their massive size, even inference for large,
> highly-accurate GPT models may require multiple performant GPUs, which
> limits the usability of such models. While there is emerging work on
> relieving this pressure via model compression, the applicability and
> performance of existing compression techniques is limited by the scale and
> complexity of GPT models. In this paper, we address this challenge, and
> propose GPTQ, a new one-shot weight quantization method based on approximate
> second-order information, that is both highly-accurate and highly-efficient.
> Specifically, GPTQ can quantize GPT models with 175 billion parameters in
> approximately four GPU hours, reducing the bitwidth down to 3 or 4 bits per

> weight, with negligible accuracy degradation relative to the uncompressed
> baseline. Our method more than doubles the compression gains relative to
> previously-proposed one-shot quantization methods, preserving accuracy,
> allowing us for the first time to execute an 175 billion-parameter model
> inside a single GPU for generative inference. Moreover, we also show that
> our method can still provide reasonable accuracy in the extreme quantization
> regime, in which weights are quantized to 2-bit or even ternary quantization
> levels. We show experimentally that these improvements can be leveraged for
> end-to-end inference speedups over FP16, of around 3.25x when using high-end
> GPUs (NVIDIA A100) and 4.5x when using more cost-effective ones (NVIDIA
> A6000). The implementation is available at [this https
> URL](https://github.com/IST-DASLab/gptq).

Comments:	ICLR 2023
Subjects:	Machine Learning (cs.LG)
Cite as:	[arXiv:2210.17323](https://arxiv.org/abs/2210.17323) [cs.LG]
(or [arXiv:2210.17323v2](https://arxiv.org/abs/2210.17323v2) [cs.LG] for this version)
<https://doi.org/10.48550/arXiv.2210.17323> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Elias Frantar [[view email](/show-email/e145ca9f/2210.17323)]
[[v1]](/abs/2210.17323v1) Mon, 31 Oct 2022 13:42:40 UTC (154 KB)
[v2] Wed, 22 Mar 2023 13:10:47 UTC (189 KB)

Full-text links:

Access Paper:

View a PDF of the paper titled GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, by Elias Frantar and 3 other authors

* [View PDF](/pdf/2210.17323)
* [TeX Source](/src/2210.17323)
* [Other Formats](/format/2210.17323)

Current browse context:

cs.LG

[< prev](/prevnext?id=2210.17323&function=prev&context=cs.LG "previous in cs.LG \(accesskey p\)") | [next >](/prevnext?id=2210.17323&function=next&context=cs.LG "next in cs.LG \(accesskey n\)")

[new](/list/cs.LG/new) | [recent](/list/cs.LG/recent) | [2022-10](/list/cs.LG/2022-10)

Change to browse by:

[cs](/abs/2210.17323?context=cs)

References & Citations

* [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2210.17323)
* [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2210.17323)
* [Semantic Scholar](https://api.semanticscholar.org/arXiv:2210.17323)

[2 blog links](/tb/2210.17323)

([what is this?](https://info.arxiv.org/help/trackback.html))

[a](/static/browse/0.3.4/css/cite.css) export BibTeX citation Loading...

BibTeX formatted citation

×

loading...

Data provided by:

Bookmark

[![BibSonomy logo](/static/browse/0.3.4/images/icons/social/bibsonomy.png)
](http://www.bibsonomy.org/BibtexHandler?requTask=upload&url=https://arxiv.org/abs/2210.17323&
description=GPTQ:
Accurate Post-Training Quantization for Generative Pre-trained Transformers

"Bookmark on BibSonomy") [![Reddit

logo](/static/browse/0.3.4/images/icons/social/reddit.png)

](https://reddit.com/submit?url=https://arxiv.org/abs/2210.17323&title=GPTQ:

Accurate Post-Training Quantization for Generative Pre-trained Transformers

"Bookmark on Reddit")

Bibliographic Tools

Bibliographic and Citation Tools

Bibliographic Explorer Toggle

Bibliographic Explorer _([What is the

Explorer?](https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-

explorer))_

Connected Papers Toggle

Connected Papers _([What is Connected

Papers?](https://www.connectedpapers.com/about))_

Litmaps Toggle

Litmaps _([What is Litmaps?](https://www.litmaps.co/))_

scite.ai Toggle

scite Smart Citations _([What are Smart Citations?](https://www.scite.ai/))_

Code, Data, Media

Code, Data and Media Associated with this Article

alphaXiv Toggle

alphaXiv _([What is alphaXiv?](https://alphaxiv.org/))_

Links to Code Toggle

CatalyzeX Code Finder for Papers _([What is CatalyzeX?](https://www.catalyzex.com))_

DagsHub Toggle

DagsHub _([What is DagsHub?](https://dagshub.com/))_

GotitPub Toggle

Gotit.pub _([What is GotitPub?](http://gotit.pub/faq))_

Huggingface Toggle

Hugging Face _([What is Huggingface?](https://huggingface.co/huggingface))_

Links to Code Toggle

Papers with Code _([What is Papers with Code?](https://paperswithcode.com/))_

ScienceCast Toggle

ScienceCast _([What is ScienceCast?](https://sciencecast.org/welcome))_

Demos

Demos

Replicate Toggle

Replicate _([What is Replicate?](https://replicate.com/docs/arxiv/about))_

Spaces Toggle

Hugging Face Spaces _([What is

Spaces?](https://huggingface.co/docs/hub/spaces))_

Spaces Toggle

TXYZ.AI _([What is TXYZ.AI?](https://txyz.ai))_

Related Papers

Recommenders and Search Tools

Link to Influence Flower

Influence Flower _([What are Influence

Flowers?](https://influencemap.cmlab.dev/))_

Core recommender toggle

CORE Recommender _([What is CORE?](https://core.ac.uk/services/recommender))_

IArxiv recommender toggle

IArxiv Recommender _([What is IArxiv?](https://iarxiv.org/about))_

* Author
* Venue
* Institution
* Topic

About arXivLabs

arXivLabs: experimental projects with community collaborators

arXivLabs is a framework that allows collaborators to develop and share new

arXiv features directly on our website.

Both individuals and organizations that work with arXivLabs have embraced and accepted our values of openness, community, excellence, and user data privacy. arXiv is committed to these values and only works with partners that adhere to them.

Have an idea for a project that will add value for arXiv's community? [**Learn more about arXivLabs**](https://info.arxiv.org/labs/index.html).

[Which authors of this paper are endorsers?](/auth/show-endorsers/2210.17323) | [Disable MathJax](javascript:setMathjaxCookie\(\)) ([What is MathJax?](https://info.arxiv.org/help/mathjax.html))

* [About](https://info.arxiv.org/about)
* [Help](https://info.arxiv.org/help)

* contact arXivClick here to contact arXiv [Contact](https://info.arxiv.org/help/contact.html)
* subscribe to arXiv mailingsClick here to subscribe [Subscribe](https://info.arxiv.org/help/subscribe)

* [Copyright](https://info.arxiv.org/help/license/index.html)
* [Privacy Policy](https://info.arxiv.org/help/policies/privacy_policy.html)

* [Web Accessibility Assistance](https://info.arxiv.org/help/web_accessibility.html)
* [arXiv Operational Status](https://status.arxiv.org)

Get status notifications via [email](https://subscribe.sorryapp.com/24846f03/email/new) or [slack](https://subscribe.sorryapp.com/24846f03/slack/new)