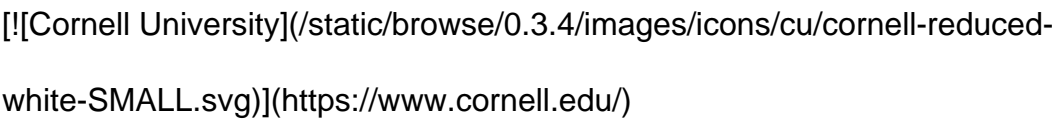



Skip to main content

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

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)


 (/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg)) (/)
> [cs] (/list/cs/recent) > arXiv:2106.09685v1

[Help] (https://info.arxiv.org/help) | [Advanced Search] (https://arxiv.org/search/advanced)

All fields Title Author Abstract Comments Journal reference ACM classification
MSC classification Report number arXiv identifier DOI ORCID arXiv author ID
Help pages Full text

Search

 (/static/browse/0.3.4/images/arxiv-logomark-small-white.svg)) (https://arxiv.org/)

[ (/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 > Computation and Language

****arXiv:2106.09685v1**** (cs)

[Submitted on 17 Jun 2021 (this version), _latest version 16 Oct 2021_
([v2](https://arxiv.org/abs/2106.09685v2))]

Title:LoRA: Low-Rank Adaptation of Large Language Models

Authors:[Edward J.

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

Shen](https://arxiv.org/search/cs?searchtype=author&query=Shen,+Y), [Phillip

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

[Zeyuan Allen-Zhu](https://arxiv.org/search/cs?searchtype=author&query=Allen-

Zhu,+Z), [Yuanzhi

Li](<https://arxiv.org/search/cs?searchtype=author&query=Li,+Y>), [Shean

Wang](<https://arxiv.org/search/cs?searchtype=author&query=Wang,+S>), [Weizhu

Chen](<https://arxiv.org/search/cs?searchtype=author&query=Chen,+W>)

View a PDF of the paper titled LoRA: Low-Rank Adaptation of Large Language Models, by Edward J. Hu and 6 other authors

[View PDF](/pdf/2106.09685v1)

> Abstract: The dominant paradigm of natural language processing consists of
> large-scale pre-training on general domain data and adaptation to particular
> tasks or domains. As we pre-train larger models, conventional fine-tuning,
> which retrains all model parameters, becomes less feasible. Using GPT-3 175B
> as an example, deploying many independent instances of fine-tuned models,
> each with 175B parameters, is extremely expensive. We propose Low-Rank
> Adaptation, or LoRA, which freezes the pre-trained model weights and injects
> trainable rank decomposition matrices into each layer of the Transformer
> architecture, greatly reducing the number of trainable parameters for
> downstream tasks. For GPT-3, LoRA can reduce the number of trainable
> parameters by 10,000 times and the computation hardware requirement by 3
> times compared to full fine-tuning. LoRA performs on-par or better than
> fine-tuning in model quality on both GPT-3 and GPT-2, despite having fewer
> trainable parameters, a higher training throughput, and no additional
> inference latency. We also provide an empirical investigation into rank-
> deficiency in language model adaptations, which sheds light on the efficacy
> of LoRA. We release our implementation in GPT-2 at [this [https](https://github.com/microsoft/LoRA)

> URL](<https://github.com/microsoft/LoRA>) .

Subjects: | Computation and Language (cs.CL); Artificial Intelligence (cs.AI); Machine Learning (cs.LG)

---|---

Cite as: | [arXiv:2106.09685](<https://arxiv.org/abs/2106.09685>) [cs.CL]

| (or [arXiv:2106.09685v1](<https://arxiv.org/abs/2106.09685v1>) [cs.CL] for this version)

| <<https://doi.org/10.48550/arXiv.2106.09685>> Focus to learn more arXiv-issued DOI via DataCite

Submission history

From: Edward J. Hu [[view email]](/show-email/e4479443/2106.09685)]

[v1] Thu, 17 Jun 2021 17:37:18 UTC (1,791 KB)

[[v2]](/abs/2106.09685v2) Sat, 16 Oct 2021 18:40:34 UTC (896 KB)

Full-text links:

Access Paper:

View a PDF of the paper titled LoRA: Low-Rank Adaptation of Large Language Models, by Edward J. Hu and 6 other authors

* [View PDF](/pdf/2106.09685v1)

* [Other Formats](/format/2106.09685v1)

[view license](<http://arxiv.org/licenses/nonexclusive-distrib/1.0/> "Rights to

this article")

Current browse context:

cs.CL

[< prev](/prevnext?id=2106.09685&function=prev&context=cs.CL "previous in cs.CL \(\accesskey p\)") | [next >](/prevnext?id=2106.09685&function=next&context=cs.CL "next in cs.CL \(\accesskey n\)")

[new](/list/cs.CL/new) | [recent](/list/cs.CL/recent) | [2021-06](/list/cs.CL/2021-06)

Change to browse by:

[cs](/abs/2106.09685?context=cs)

[cs.AI](/abs/2106.09685?context=cs.AI)

[cs.LG](/abs/2106.09685?context=cs.LG)

References & Citations

- * [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2106.09685)
- * [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2106.09685)
- * [Semantic Scholar](https://api.semanticscholar.org/arXiv:2106.09685)

[12 blog links](/tb/2106.09685)

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

[DBLP](https://dblp.uni-trier.de) \- CS Bibliography

[listing](https://dblp.uni-trier.de/db/journals/corr/corr2106.html#abs-2106-09685 "listing on DBLP") |

[bibtex](https://dblp.uni-trier.de/rec/bibtex/journals/corr/abs-2106-09685 "DBLP bibtex record")

[Yelong Shen](https://dblp.uni-trier.de/search/author?author=Yelong%20Shen
"DBLP author search")

[Phillip Wallis](https://dblp.uni-
trier.de/search/author?author=Phillip%20Wallis "DBLP author search")

[Zeyuan Allen-Zhu](https://dblp.uni-
trier.de/search/author?author=Zeyuan%20Allen-Zhu "DBLP author search")

[Yuanzhi Li](https://dblp.uni-trier.de/search/author?author=Yuanzhi%20Li "DBLP
author search")

[Weizhu Chen](https://dblp.uni-trier.de/search/author?author=Weizhu%20Chen
"DBLP author search")

[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/2106.09685&](http://www.bibsonomy.org/BibtexHandler?requTask=upload&url=https://arxiv.org/abs/2106.09685&description=LoRA:)
description=LoRA:

Low-Rank Adaptation of Large Language Models "Bookmark on BibSonomy") [

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

(<https://reddit.com/submit?url=https://arxiv.org/abs/2106.09685&title=LoRA:>

Low-Rank Adaptation of Large Language Models "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-](https://info.arxiv.org/labs/showcase.html#arxiv-bibliographic-explorer)
explorer))_

Connected Papers Toggle

Connected Papers _([What is Connected

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

Litmaps Toggle

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

scite.ai Toggle

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

Code, Data, Media

Code, Data and Media Associated with this Article

alphaXiv Toggle

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

Links to Code Toggle

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

DagsHub Toggle

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

GotitPub Toggle

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

XYZ.AI _([What is XYZ.AI?](https://xyz.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))_

- * 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? [\[**Learn more about arXivLabs**\]\(https://info.arxiv.org/labs/index.html\)](https://info.arxiv.org/labs/index.html).

[\[Which authors of this paper are endorsers?\]\(/auth/show-endorsers/2106.09685\)](/auth/show-endorsers/2106.09685) | [\[Disable MathJax\]\(javascript:setMathjaxCookie\\(\\)\)](#) [\(\[What is MathJax?\]\(https://info.arxiv.org/help/mathjax.html\)\)](#)

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

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

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

* [subscribe to arXiv mailingsClick here to subscribe \[Subscribe\]\(https://info.arxiv.org/help/subscribe\)](https://info.arxiv.org/help/subscribe)

* [\[Copyright\]\(https://info.arxiv.org/help/license/index.html\)](https://info.arxiv.org/help/license/index.html)

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

* [\[Web Accessibility Assistance\]\(https://info.arxiv.org/help/web_accessibility.html\)](https://info.arxiv.org/help/web_accessibility.html)

* [\[arXiv Operational Status \]\(https://status.arxiv.org\)](https://status.arxiv.org)

Get status notifications via

[\[email\]\(https://subscribe.sorryapp.com/24846f03/email/new\)](https://subscribe.sorryapp.com/24846f03/email/new) or

[\[slack\]\(https://subscribe.sorryapp.com/24846f03/slack/new\)](https://subscribe.sorryapp.com/24846f03/slack/new)

