Skip to main content

[![Cornell University](/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)

[![arxiv logo](/static/browse/0.3.4/images/arxiv-logo-one-color-white.svg)](/) > [cs](/list/cs/recent) > arXiv:2106.09650

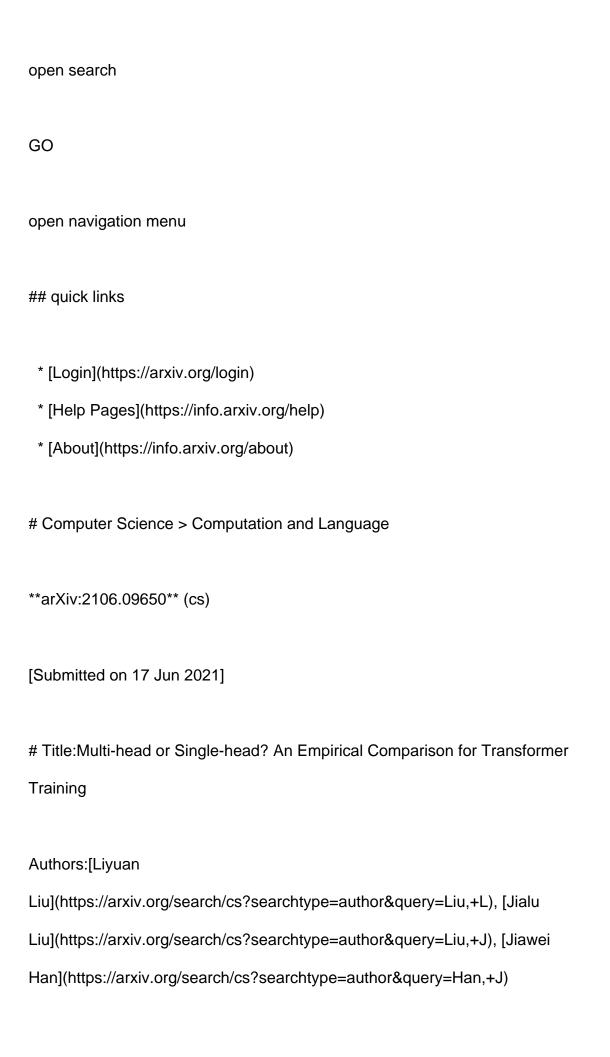
[Help](https://info.arxiv.org/help) | [Advanced Search](https://arxiv.org/search/advanced)

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

[![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/)



View a PDF of the paper titled Multi-head or Single-head? An Empirical

Comparison for Transformer Training, by Liyuan Liu and Jialu Liu and Jiawei

Han

[View PDF](/pdf/2106.09650)

> Abstract: Multi-head attention plays a crucial role in the recent success of

> Transformer models, which leads to consistent performance improvements over

> conventional attention in various applications. The popular belief is that

> this effectiveness stems from the ability of jointly attending multiple

> positions. In this paper, we first demonstrate that jointly attending

> multiple positions is not a unique feature of multi-head attention, as

> multi-layer single-head attention also attends multiple positions and is

> more effective. Then, we suggest the main advantage of the multi-head

> attention is the training stability, since it has less number of layers than

> the single-head attention, when attending the same number of positions. For

> example, 24-layer 16-head Transformer (BERT-large) and 384-layer single-head

> Transformer has the same total attention head number and roughly the same

> model size, while the multi-head one is significantly shallower. Meanwhile,

> we show that, with recent advances in deep learning, we can successfully

> stabilize the training of the 384-layer Transformer. As the training

> difficulty is no longer a bottleneck, substantially deeper single-head

> Transformer achieves consistent performance improvements without tuning

> hyper-parameters.

Comments: | Work in progress

---|---

Subjects: | Computation and Language (cs.CL); Machine Learning (cs.LG) Cite as: | [arXiv:2106.09650](https://arxiv.org/abs/2106.09650) [cs.CL] (or [arXiv:2106.09650v1](https://arxiv.org/abs/2106.09650v1) [cs.CL] for this version) | <https://doi.org/10.48550/arXiv.2106.09650> Focus to learn more arXiv-issued DOI via DataCite ## Submission history From: Liyuan Liu [[view email](/show-email/d6d762a3/2106.09650)] **[v1]** Thu, 17 Jun 2021 16:53:22 UTC (295 KB) Full-text links: ## Access Paper: View a PDF of the paper titled Multi-head or Single-head? An Empirical Comparison for Transformer Training, by Liyuan Liu and Jialu Liu and Jiawei Han * [View PDF](/pdf/2106.09650) * [TeX Source](/src/2106.09650) * [Other Formats](/format/2106.09650) [view license](http://arxiv.org/licenses/nonexclusive-distrib/1.0/ "Rights to this article") Current browse context:

```
[< prev](/prevnext?id=2106.09650&function=prev&context=cs.CL "previous in cs.CL \(accesskey p\)") | [next >](/prevnext?id=2106.09650&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/2106.09650?context=cs)

[cs.LG](/abs/2106.09650?context=cs.LG)

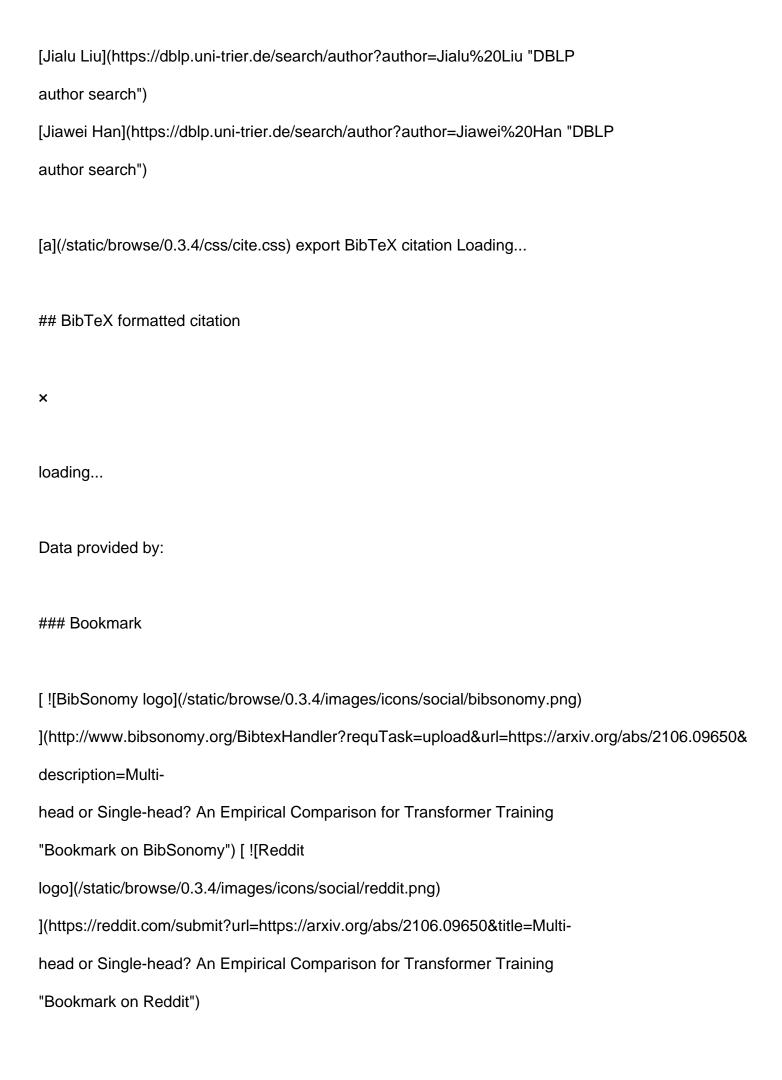
References & Citations

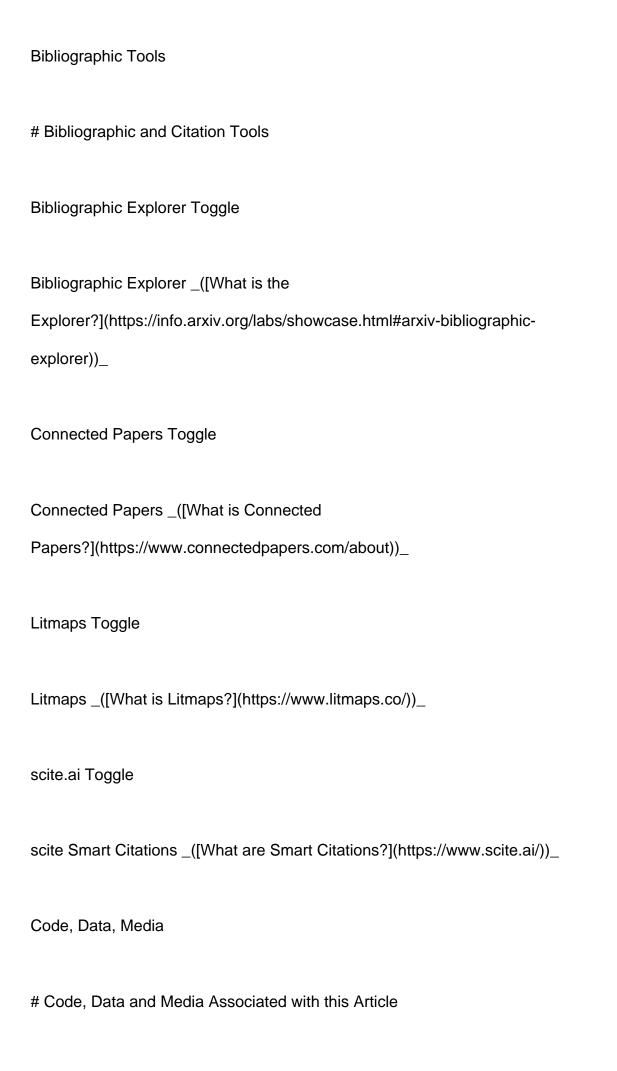
- * [NASA ADS](https://ui.adsabs.harvard.edu/abs/arXiv:2106.09650)
- * [Google Scholar](https://scholar.google.com/scholar_lookup?arxiv_id=2106.09650)
- * [Semantic Scholar](https://api.semanticscholar.org/arXiv:2106.09650)

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

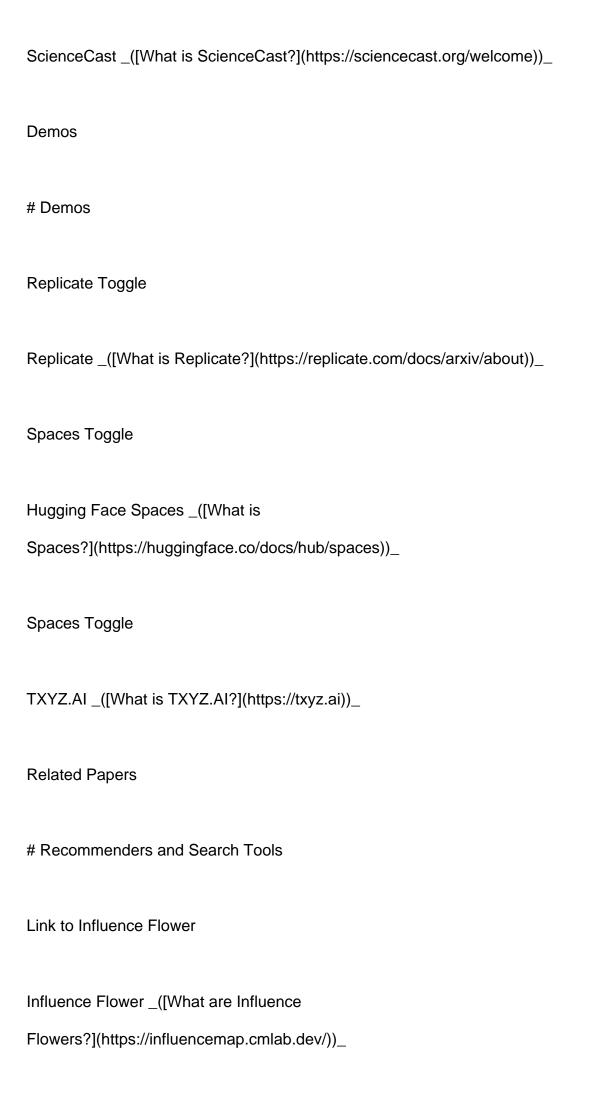
[listing](https://dblp.uni-trier.de/db/journals/corr/corr2106.html#abs-2106-09650 "listing on DBLP") | [bibtex](https://dblp.uni-trier.de/rec/bibtex/journals/corr/abs-2106-09650 "DBLP bibtex record")

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





```
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
```



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).

[Which authors of this paper are endorsers?](/auth/show-endorsers/2106.09650) | [Disable MathJax](javascript:setMathjaxCookie\(\)) ([What is

MathJax?](htt	ps://info.arxiv.o	rg/help/	mathjax.ht	ml))				
* [About](http	os://info.arxiv.or	g/about	:)					
* [Help](https	s://info.arxiv.org	/help)						
* contact arX	(ivClick here to	contact	arXiv [Co	ntact](https://info.a	ırxiv.org/h	elp/cont	act.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)								
[i iivacy i c	moy](nttps://ime	,.aixiv.o	ту/пстр/ро	iloics/privacy_polic	,y .11(1111)			
* [Web Acce	ssibility Assista	nce](htt	ps://info.ar	xiv.org/help/web_a	accessibili	ty.html)		
* [arXiv Ope	rational Status]	(https://	⁄status.arxi	v.org)				
Get status not	tifications via							

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

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