

[Skip to main content](#)> [cs](#) > arXiv:1706.03762

quick links

- [Login](#)
- [Help Pages](#)
- [About](#)

Computer Science > Computation and Language

arXiv:1706.03762 (cs)

[Submitted on 12 Jun 2017 ([v1](#)), last revised 6 Dec 2017 (this version, v5)]

Attention Is All You Need

[Ashish Vaswani](#), [Noam Shazeer](#), [Niki Parmar](#), [Jakob Uszkoreit](#), [Llion Jones](#),
[Aidan N. Gomez](#), [Lukasz Kaiser](#), [Illia Polosukhin](#)

[Download PDF](#)


The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Comments: 15 pages, 5 figures

Subjects: **Computation and Language (cs.CL)**; Machine Learning (cs.LG)Cite as: [arXiv:1706.03762](#) [cs.CL]

(or [arXiv:1706.03762v5](https://arxiv.org/abs/1706.03762v5) [cs.CL] for this version)

<https://doi.org/10.48550/arXiv.1706.03762>

 Focus to learn more

arXiv-issued DOI via DataCite

Submission history

From: Ashish Vaswani [[view email](#)]

[\[v1\]](#) Mon, 12 Jun 2017 17:57:34 UTC (1,102 KB)

[\[v2\]](#) Mon, 19 Jun 2017 16:49:45 UTC (1,125 KB)

[\[v3\]](#) Tue, 20 Jun 2017 05:20:02 UTC (1,125 KB)

[\[v4\]](#) Fri, 30 Jun 2017 17:29:30 UTC (1,124 KB)

[\[v5\]](#) Wed, 6 Dec 2017 03:30:32 UTC (1,124 KB)

☐ Bibliographic Tools

Bibliographic and Citation Tools

☐ Bibliographic Explorer Toggle

Bibliographic Explorer ([What is the Explorer?](#))

☐ Litmaps Toggle

Litmaps ([What is Litmaps?](#))

☐ scite.ai Toggle

scite Smart Citations ([What are Smart Citations?](#))

☒ Code, Data, Media

Code, Data and Media Associated with this Article

☐ Links to Code Toggle

CatalyzeX Code Finder for Papers ([What is CatalyzeX?](#))

☐ DagsHub Toggle

DagsHub ([What is DagsHub?](#))

☐ Links to Code Toggle

Papers with Code ([What is Papers with Code?](#))

☐ ScienceCast Toggle

ScienceCast ([What is ScienceCast?](#))

☐ Demos

Demos

☐ Replicate Toggle

Replicate ([What is Replicate?](#))

☐ Spaces Toggle

Hugging Face Spaces ([What is Spaces?](#))

☐ Related Papers

Recommenders and Search Tools

☐ Link to Influence Flower

Influence Flower ([What are Influence Flowers?](#))

☐ Connected Papers Toggle

Connected Papers ([What is Connected Papers?](#))

☐ Core recommender toggle

CORE Recommender ([What is CORE?](#))

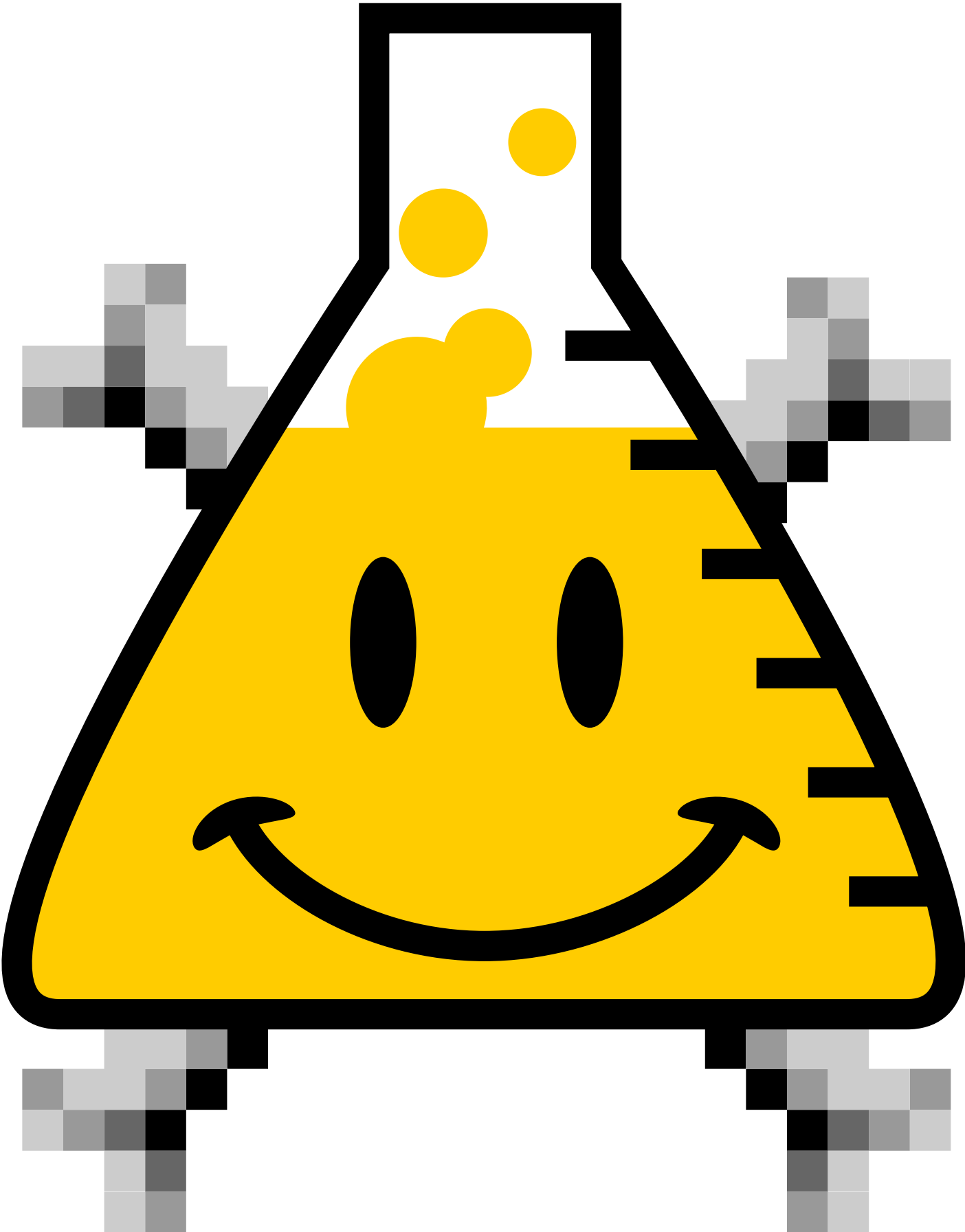
☐ About arXivLabs

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](#).



[Which authors of this paper are endorsers?](#) | [Disable MathJax](#) ([What is MathJax?](#))