
Convolutions for Sequence Modeling

This repository provides implementations and experiments for the following papers, as well as simplified presentations of earlier work such as S4.

Please see these instructions for how to download weights and run our pretrained models: * H3 (125m-2.7B) * Hyena (small, 150M)

Hyena

Hyena Hierarchy: Towards Larger Convolutional Language models Michael Poli*, Stefano Massaroli*, Eric Nguyen*, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, Christopher Ré
ICML 2023. **Oral.**

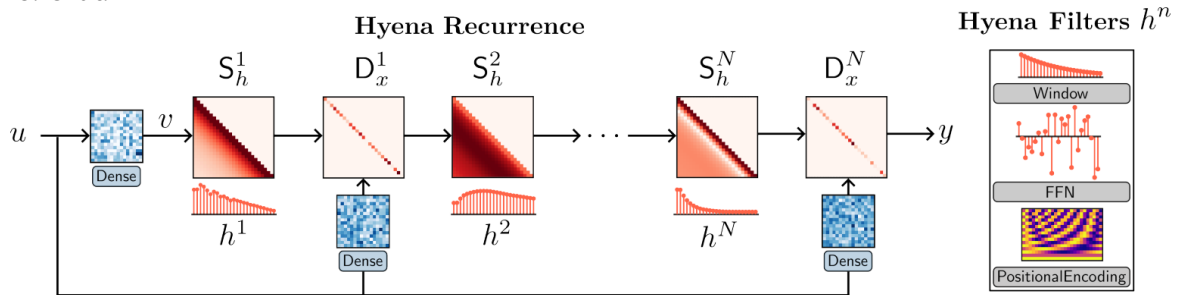


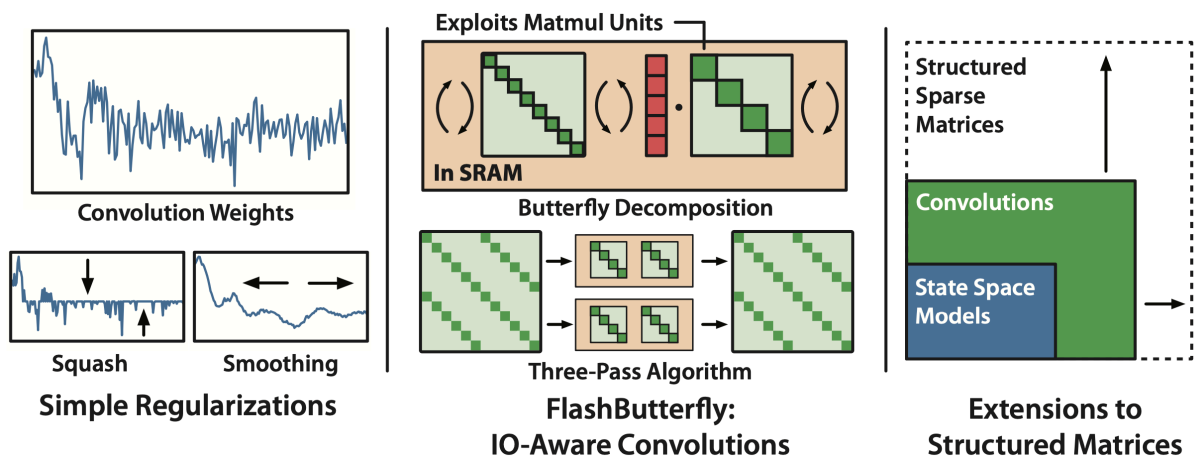
Figure 1.1: The Hyena operator is defined as a recurrence of two efficient subquadratic primitives: an implicit long convolution h (i.e. Hyena filters parameterized by a feed-forward network) and multiplicative element-wise gating of the (projected) input. The depth of the recurrence specifies the size of the operator. Hyena can equivalently be expressed as a multiplication with *data-controlled* (conditioned by the input u) diagonal matrices D_x and Toeplitz matrices S_h .

Paper

Long Convs

Simple Hardware-Efficient Long Convolutions for Sequence Modeling

Daniel Y. Fu, *Elliot L. Epstein*, Eric Nguyen, Armin W. Thomas, Michael Zhang, Tri Dao, Atri Rudra, Christopher Ré
ICML 2023.



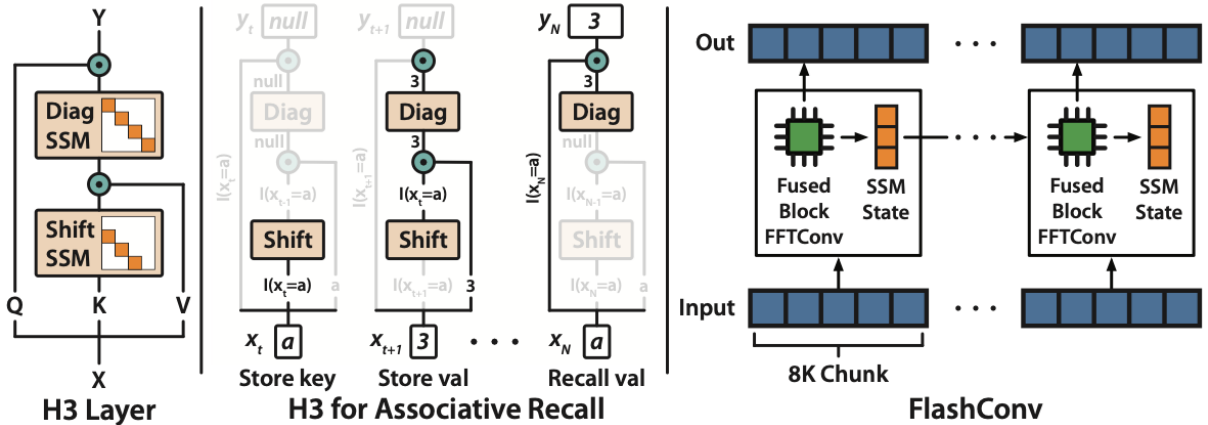
Paper

Hungry Hungry Hippos (H3)

Hungry Hungry Hippos: Towards Language Modeling with State Space Models

Daniel Y. Fu*, Tri Dao*, Khaled K. Saab, Armin W. Thomas, Atri Rudra, Christopher Ré

ICLR 2023. **Notable top-25% (spotlight).**



Paper

Roadmap

- Include H3, LLM training, and synthetics in this repository
- Move in fast convolution code
- Add Hyena implementation and experiments
- pip package

Changelog

See CHANGELOG.md

Setup

Requirements

This repository requires Python 3.8+ and Pytorch 1.10+. Other packages are listed in requirements.txt.

Getting Started

The easiest way to get started is to run the `standalone_cifar.py` script. This script trains a simple long convolution model on CIFAR-10:

```
1 python -m standalone_cifar
```

See the experiments page for more: * LRA experiments from the Long Convs paper * H3 experiments (language model, synthetics) * H3 + Long Conv experiments * Hyena language and vision experiments

Resources

We're happy to share independent reimplementations and explainer posts about methods presented in this repository.

Hyena:

- irhum's JAX reimplementation and comparison with nanoGPT
- exps's reimplementation and explainer post

Citation

If you use this codebase, or otherwise found our work valuable, you can cite us as follows:

```
1 @article{poli2023hyena,  
2   title={Hyena Hierarchy: Towards Larger Convolutional Language Models  
   },
```

```
3   author={Poli, Michael and Massaroli, Stefano and Nguyen, Eric and Fu,
      Daniel Y and Dao, Tri and Baccus, Stephen and Bengio, Yoshua and
      Ermon, Stefano and R{\e}, Christopher},
4   journal={arXiv preprint arXiv:2302.10866},
5   year={2023}
6 }
7
8 @article{fu2023simple,
9   title={Simple Hardware-Efficient Long Convolutions for Sequence
      Modeling},
10  author={Fu, Daniel Y. and Epstein, Elliot L. and Nguyen, Eric and
      Thomas, Armin W. and Zhang, Michael and Dao, Tri and Rudra, Atri
      and R{\e}, Christopher},
11  journal={International Conference on Machine Learning},
12  year={2023}
13 }
14
15 @inproceedings{fu2023hungry,
16   title={Hungry {H}ungry {H}ippos: Towards Language Modeling with State
      Space Models},
17   author={Fu, Daniel Y. and Dao, Tri and Saab, Khaled K. and Thomas,
      Armin W.
18   and Rudra, Atri and R{\e}, Christopher},
19   booktitle={International Conference on Learning Representations},
20   year={2023}
21 }
```

Acknowledgements

This repo was forked from Albert Gu's state spaces repo and borrows its structure. It also contains code from the FlashAttention training scripts.