Transformer Quality in Linear Time

Weizhe Hua[1,2], *Zihang Dai*[*2], *Hanxiao Liu*[*2], *Quoc V. Le*[2]
*Equal contribution
Cornell Univeristy[1]; Google Research, Brain Team[2]

Session 8 Track 3 @ICML 2022

Google Research

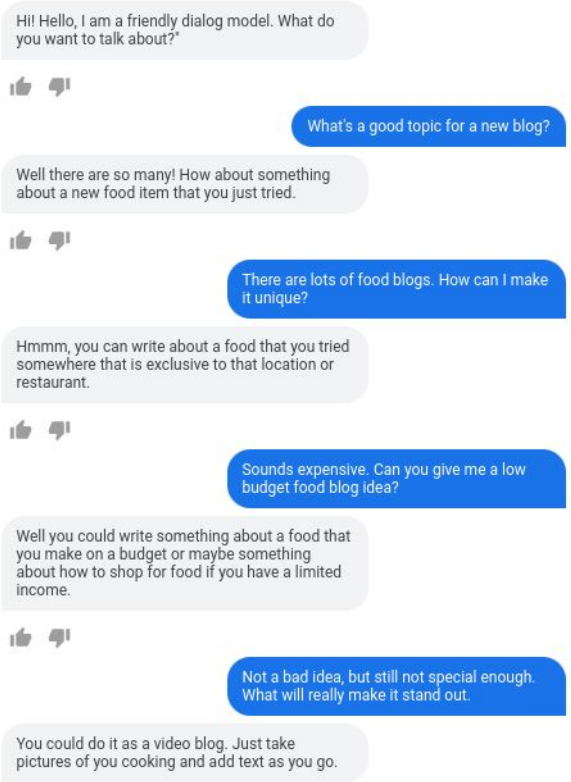
Motivation

Emerging applications need to handle long contexts.

Transformers: $O(T^2)$ over context length T.

O(T) variants exist, but are rarely deployed:

- Inferior Quality
- Overhead in Practice
  - Memory re-formatting ops [Tay ICLR'21]
- Inefficient Auto-regressive Training
  - Sequential state updates in linear attention Q(ϕ(K)^Tϕ(V)) [Choromanski ICLR'21]



Google LaMDA

credit: ai.googleblog.com

Contribution

Key Ideas

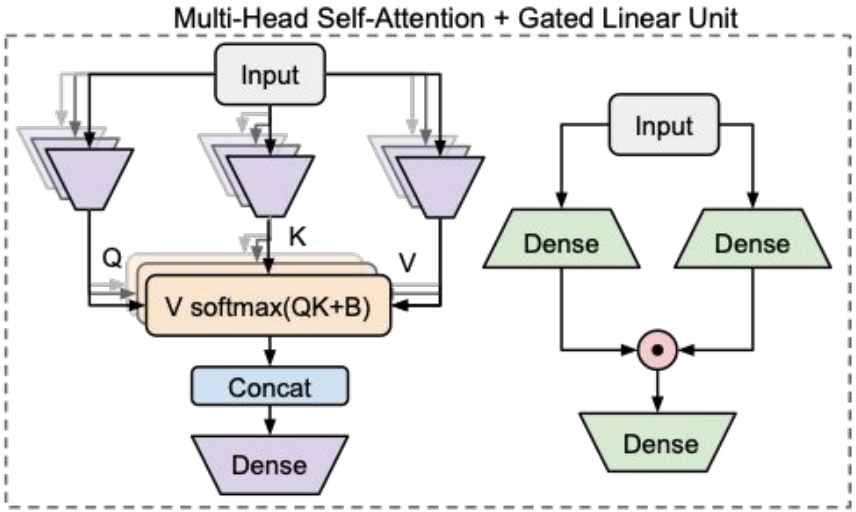
- Simplify the Transformer layer to make it more suitable for approximation
- An accelerator-friendly approximation of attention

Our model — ***FLASH***

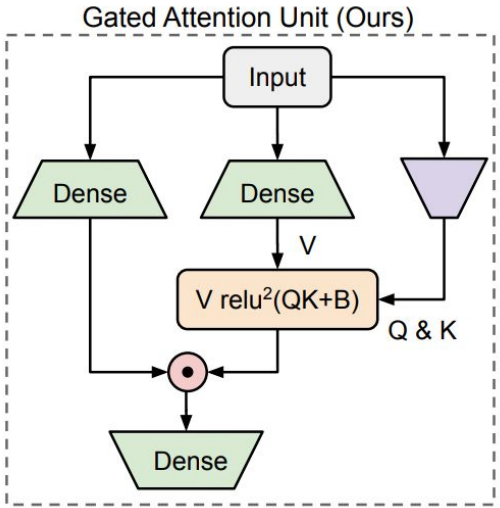
***F**ast **L**inear **A**ttention with a **S**ingle **H**ead:*

- **Quality**: ~= Augmented Transformer
- **Complexity**: $O(T^2) \Rightarrow O(T)$
- **Speed**: Fast on TPUs & GPUs

Replace Transformer with Gated Attention Unit

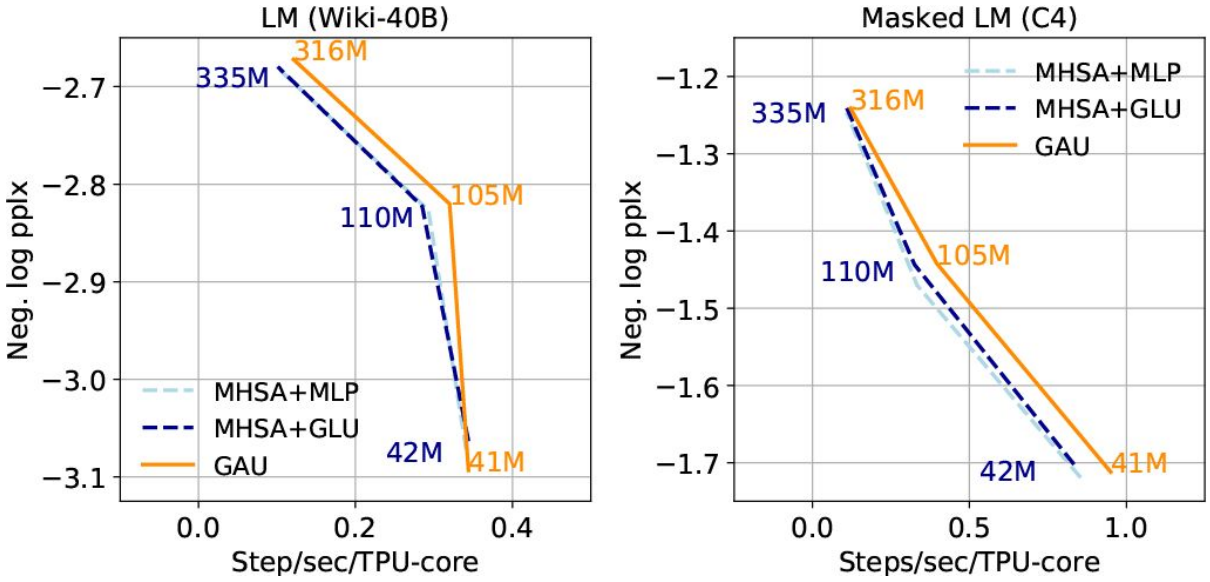


Most ops in MHSA (**purple**) are removed



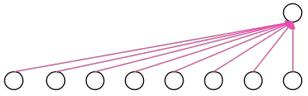
- **Compute & Memory Savings:**
  - Shared *Q* and *K* with small hidden dimension (e.g., 128)
  - Single head attention reduces memory complexity from $O(HT^2)$ to $O(T^2)$
- **Friendly to further approximation of the attention:**
  - Enable deeper attention stacks: Transformer layer ⇒ 2 GAUs

GAU vs. Transformers (short seqs)

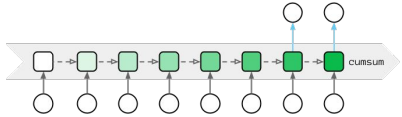


*GAUs are **at least competitive** with Transformers across different model sizes even when the context length is relatively short (512).*

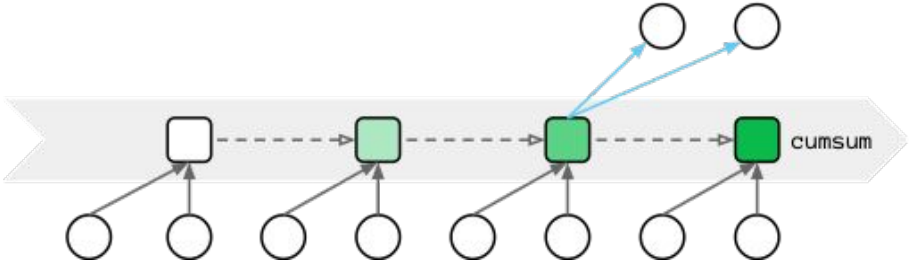
Mixed Chunk Attention (ours)



quadratic



linear



Ours

Reduced Compute:

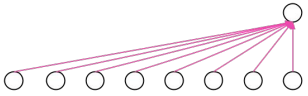
global linear (→)

$$o_i = q_i \underbrace{\left(\sum_{j \in \text{global}} \mathbf{k_j} \mathbf{v_j}^T \right)}_{\text{linear}}$$

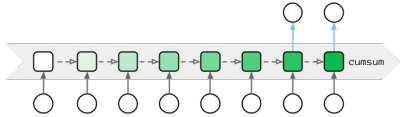
Pseudocode (**n**:# of chunks, **c**:chunk size)

```
lin_kv = enisum('bncd,bnce->bnde', k, v)
lin_kv = cumsum(lin_kv, axis=1, exclusive=True)
lin = enisum('bncd,bnde->bnce',q, lin_kv)
```

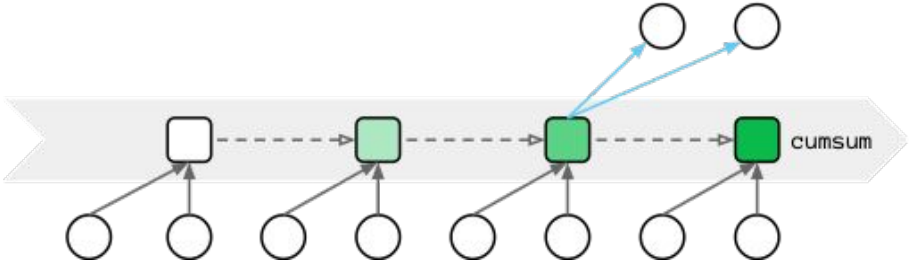
Mixed Chunk Attention (ours)



quadratic



linear



Ours

Reduced Compute:

global linear (→)

$$o_i = q_i \underbrace{\left(\sum_{j \in \text{global}} \mathbf{k_j} \mathbf{v_j}^T \right)}_{\text{linear}}$$

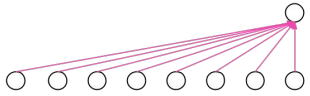
Pseudocode (**n**:# of chunks, **c**:chunk size)

```
lin_kv = enisum('bncd,bnce->bnde', k, v)
lin_kv = cumsum(lin_kv, axis=1, exclusive=True)
lin = enisum('bncd,bnde->bnce',q, lin_kv)
```

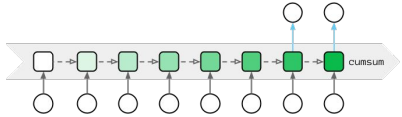
Reduced RNN-like steps/states (■)

# of tokens (e.g., 8K) ⇒ # of states (e.g., 16)

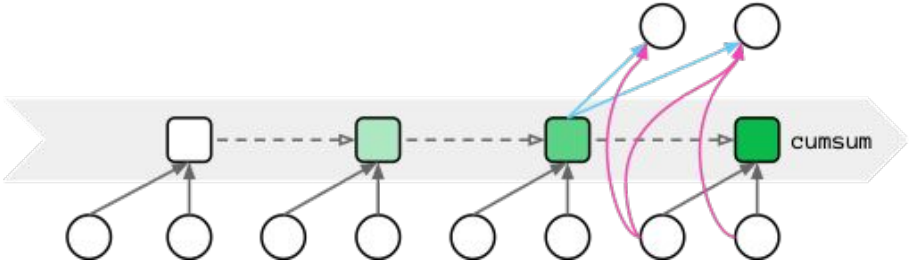
Mixed Chunk Attention (ours)



quadratic



linear



Ours

Reduced Compute:

global linear (→) + Local quadratic (→)

$$o_i = q_i \underbrace{\left(\sum_{j \in \text{global}} \mathbf{k_j}\, \mathbf{v_j}^T \right)}_{\text{linear}} + \underbrace{\sum_{j \in \text{local}} \text{relu}^2\left(\mathbf{q_i}^T \mathbf{k_j} \right)}_{\text{quadratic}} v_j$$

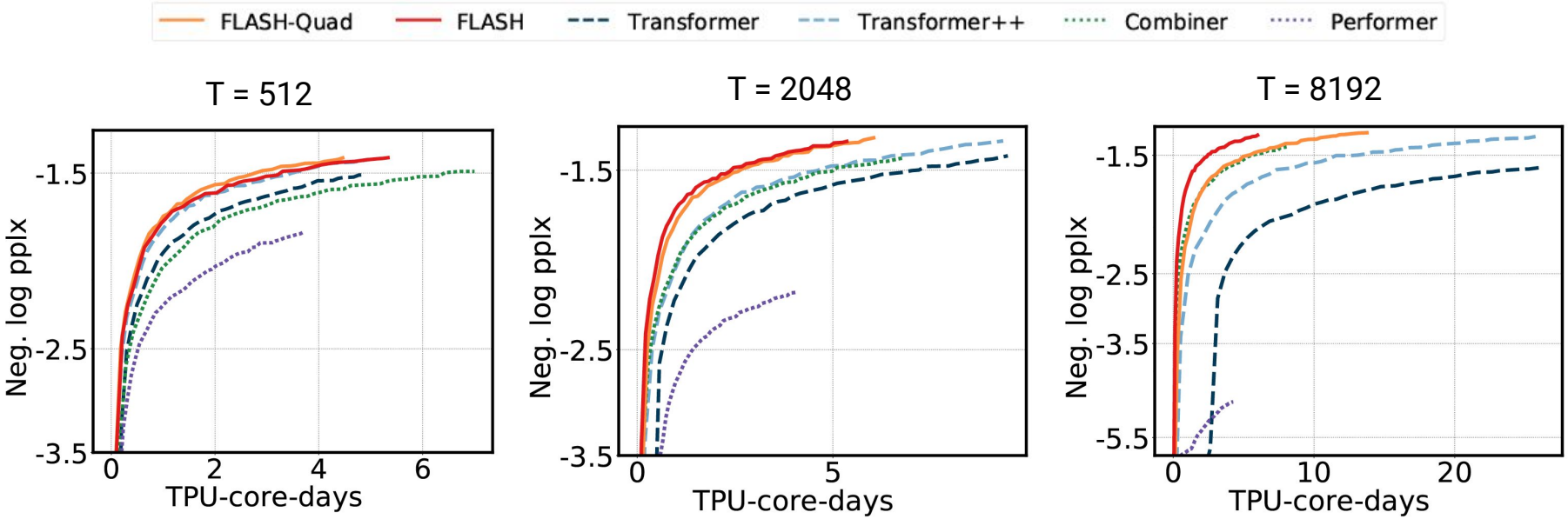
Pseudocode (**n**:# of chunks, **c**:chunk size)

```
lin_kv = enisum('bncd,bnce->bnde', k, v)
lin_kv = cumsum(lin_kv, axis=1, exclusive=True)
lin = enisum('bncd,bnde->bnce',q, lin_kv)
```

Reduced RNN-like steps/states (■)

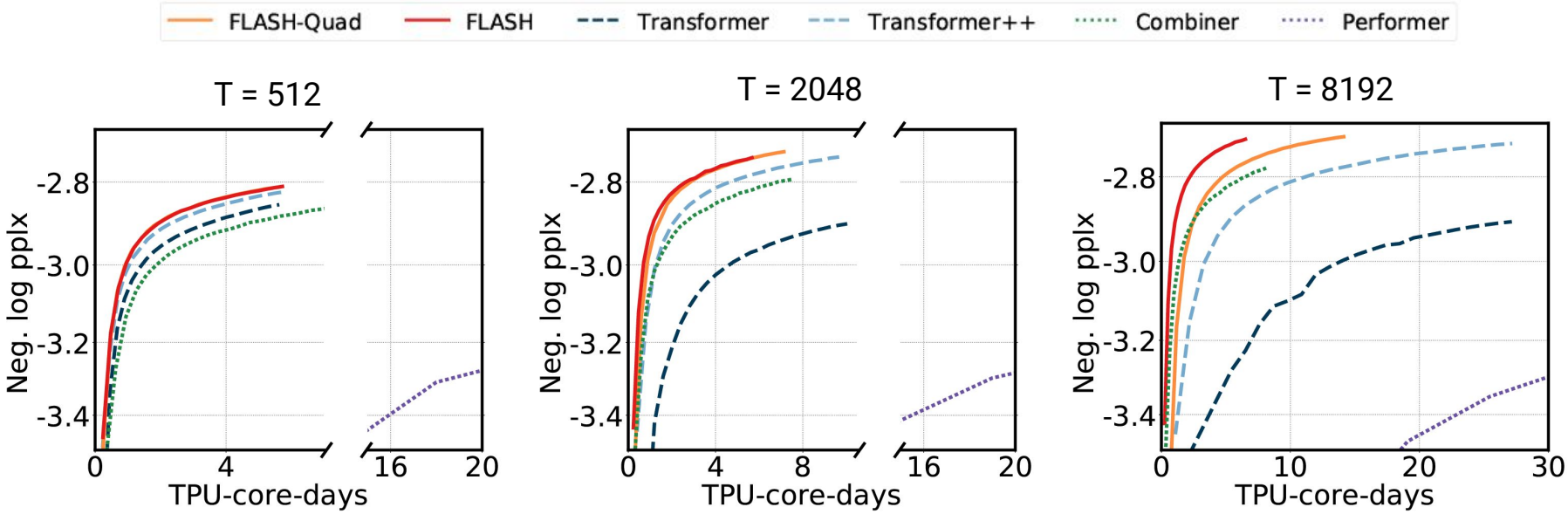
# of tokens (e.g., 8K) ⇒ # of states (e.g., 16)

Masked Language Modeling on C4



*FLASH reduces the training cost of Transformer++ by 1.0× (T=512) – **4.8× (T=8K)**.*

Language Modeling on Wiki-40B



*FLASH reduces the training cost of Transformer++ by 1.2× (T=512) – **4.9× (T=8K)**.*

Language Modeling on PG-19

PG-19 [Rae ICLR'20]: books extracted from Project Gutenberg

- Avg. length: 69K words

Model	Context Length											
	1024			2048			4096			8192		
	PPLX	Lat.	Speedup*	PPLX	Lat.	Speedup*	PPLX	Lat.	Speedup*	PPLX	Lat.	Speedup*
Transformer+	44.45	282	1.00×	43.14	433	1.00×	42.80	698	1.00×	43.27	1292	1.00×
Transformer++	44.47	292	–	43.18	441	–	43.13	712	–	43.26	1272	1.21×
Combiner	46.04	386	–	44.68	376	–	43.99	374	–	44.12	407	–
FLASH-Quad	**43.40**	**231**	**2.18×**	**42.01**	273	3.29×	41.46	371	3.59×	41.68	560	5.23×
FLASH	44.06	**234**	1.66×	42.17	**237**	**3.85×**	**40.72**	**234**	**6.75×**	**41.07**	**250**	**12.12×**

*FLASH outperforms the quality of Transformer+ while being up to **12.1x** faster!*

Fine-tuning on TriviaQA

TriviaQA [Joshi ACL'17]: Passages can span multiple documents

- Pretraining and fine-tuning with 4K context length

Model	PT PPLX	FT F1	PT / FT Lat. reduction
Transformer+	3.48	74.2	1.00× / 1.00×
Combiner	3.51	67.2	**2.78× / 2.75×**
FLASH-Quad$_{s=128}$	3.24	72.7	1.89× / 1.79×
FLASH-Quad$_{s=512}$	**3.12**	**74.8**	1.76× / 1.67×
FLASH$_{s=512}$	3.23	73.3	2.61× / 2.60×
FLASH$_{s=512}$ + first-to-all	3.24	73.9	**2.78× / 2.69×**

*FLASH reduces the pretraining/fine-tuning cost of Transformer+ by **2.8×** and **2.7×**.*

Summary

More in the paper:

- Experimental setups & implementation details
- Ablation studies

Conclusion:

- We propose a high-quality and high-performance sequence model, called **FLASH**, which achieves the **transformer quality in linear time**.

Thank you!

Transformer Quality in Linear Time

Weizhe Hua[12], *Zihang Dai*[2], *Hanxiao Liu*[2], *Quoc Le*[2]

*Equal contribution

Cornell Univeristy[1]; Google Research, Brain Team[2]

Session 8 Track 3 @ICML 2022

Google Research