Dissecting GPT: The Complete Forward Pass

Jacob Danner

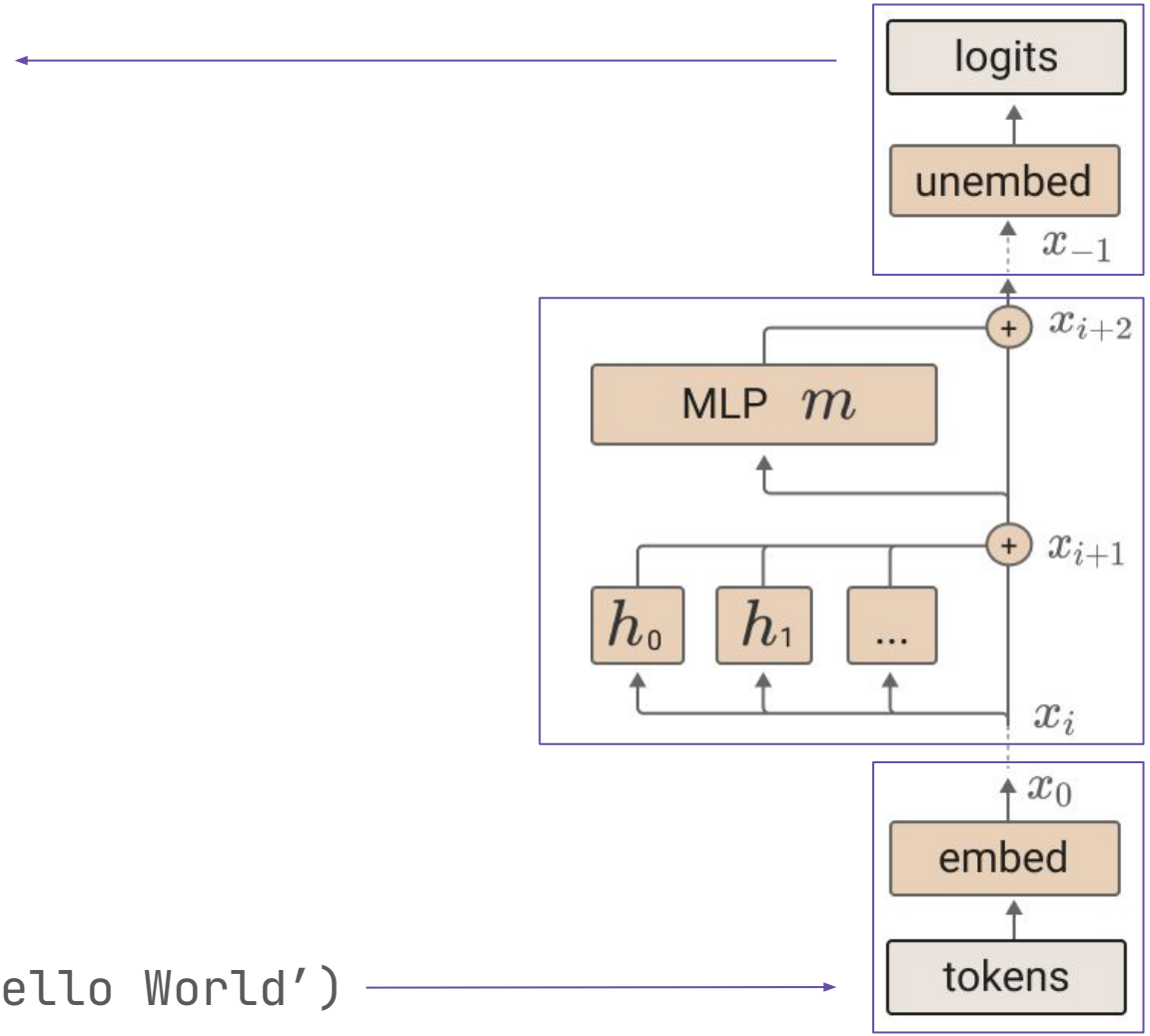
Language Modeling

- *“Attention Is All You Need”* (2017, Google)
  - **Innovation**: The Transformer architecture.
  - **Breakthrough**: General purpose algorithm for processing data in parallel using multi-head self-attention, not sequentially
  - **Impact**: Enabled larger models and faster training on GPUs
- *“Improving Language Understanding by Generative Pre-Training”* (2018, OpenAI)
  - **Innovation**: Generative Pre-Training
  - **Breakthrough**: Trained a model on unsupervised text to learn universal language skills
  - **Result**: Created a single model foundation that could be fine-tuned for a wide variety of tasks

GPT: What to Know

- Input
  - Tokens
  - Embed (token + positional)
- Processing (transformer blocks)
  - Multi-Head Self-Attention
  - MLP
  - Residual streams
- Output
  - Unembed
  - Sampling

.	60%
!	20%
,	8%
…	…

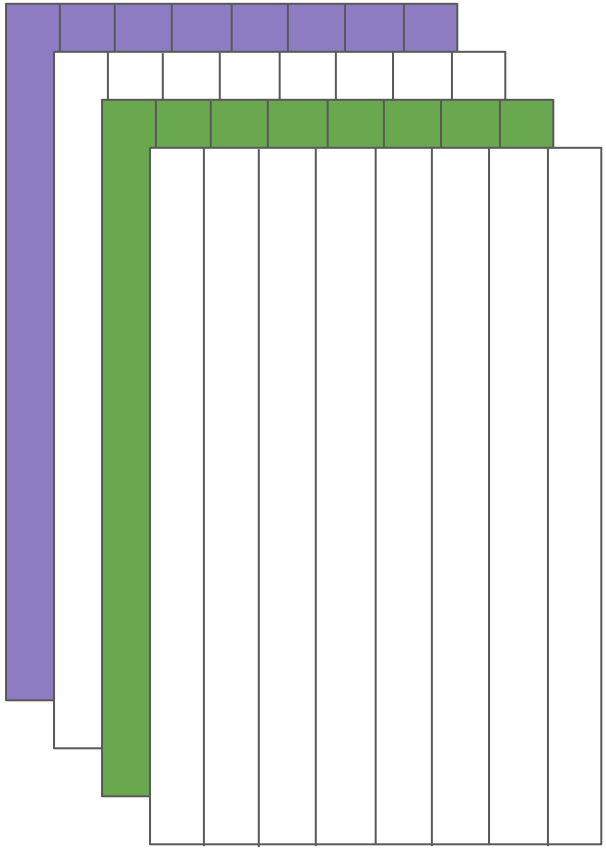


tokenizer(‘Hello World’)

$$h_0 = UW_e + W_p$$
$$h_l = \texttt{transformer\_block}(h_{l-1}) \forall i \in [1, n]$$
$$P(u) = \texttt{softmax}(h_n W_e^T)$$

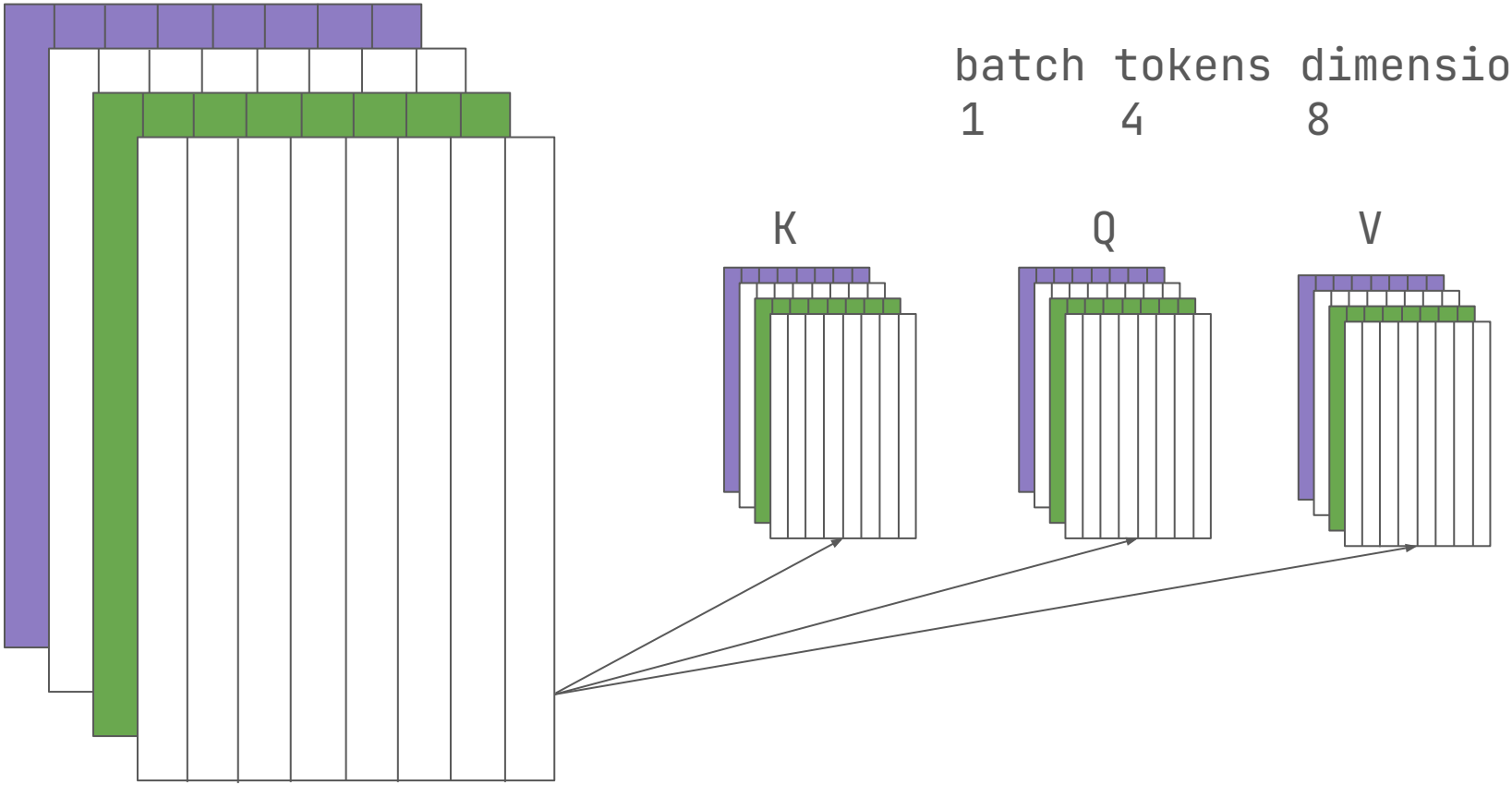
Initial Residual Streams

He
llo
Wo
rld



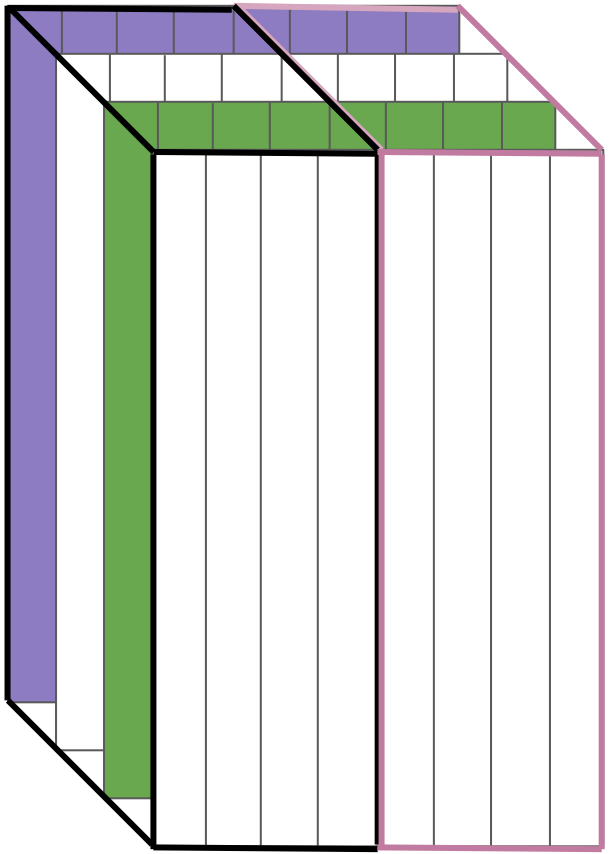
batch tokens dimension
1 4 8

Initial Residual Streams (K, Q, V)



2 attention heads

He
llo
Wo
rld



batch	n_heads	tokens	dimension
1	2	4	4

Attention (looking at 1 head)

Q @ K^T

	He	llo	Wo	rld
He	1	-	-	-
llo	.7	.3	-	-
Wo	.2	.2	.6	-
rld	.2	.2	.2	.4

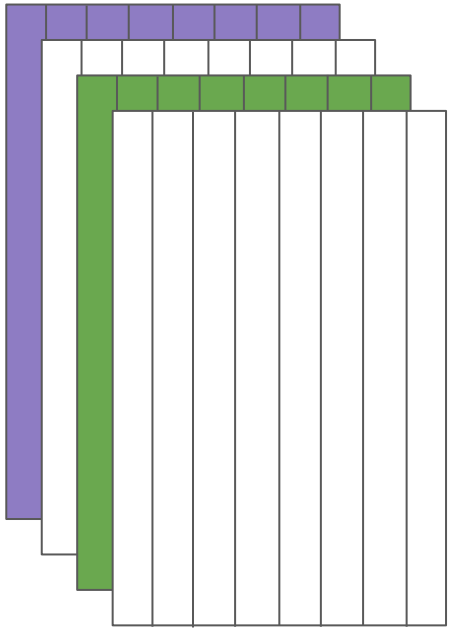
Attention (looking at 1 head)

Q @ K^T

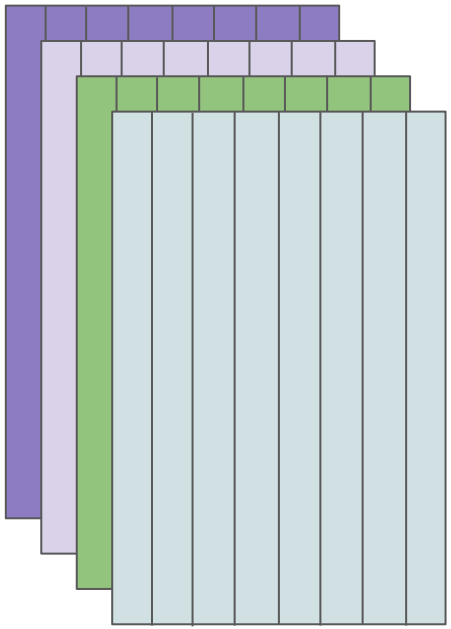
	He	llo	Wo	rld
He	1	-	-	-
llo	.7	.3	-	-
Wo	.2	.2	.6	-
rld	.2	.2	.2	.4

✖

V

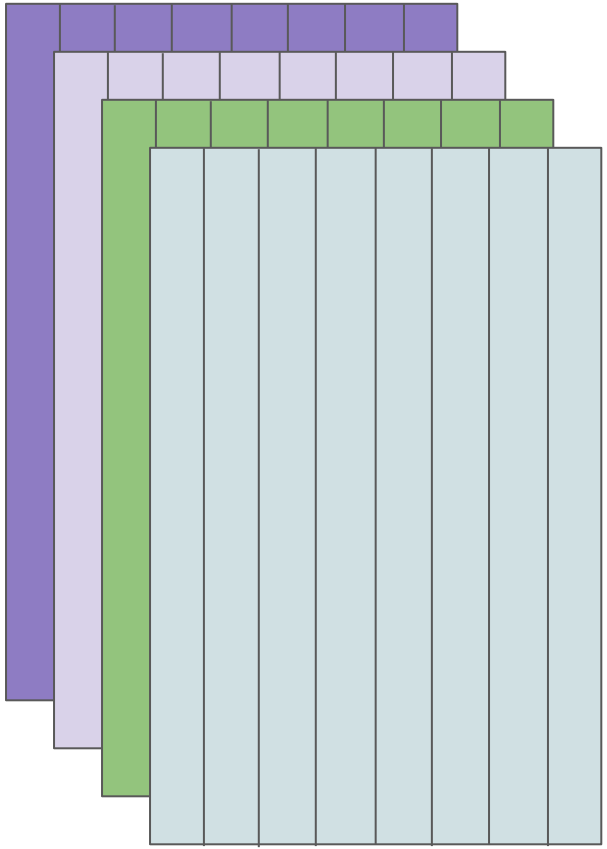


=



Residual Streams After Attention

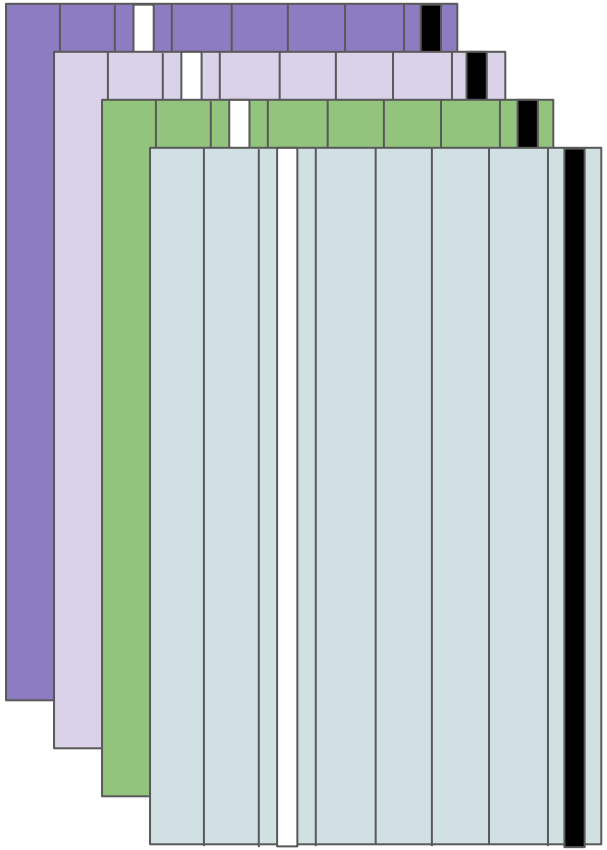
He
llo
Wo
rld



batch tokens dimension
1 4 8

Residual Streams After Attention and MLP

He
llo
Wo
rld

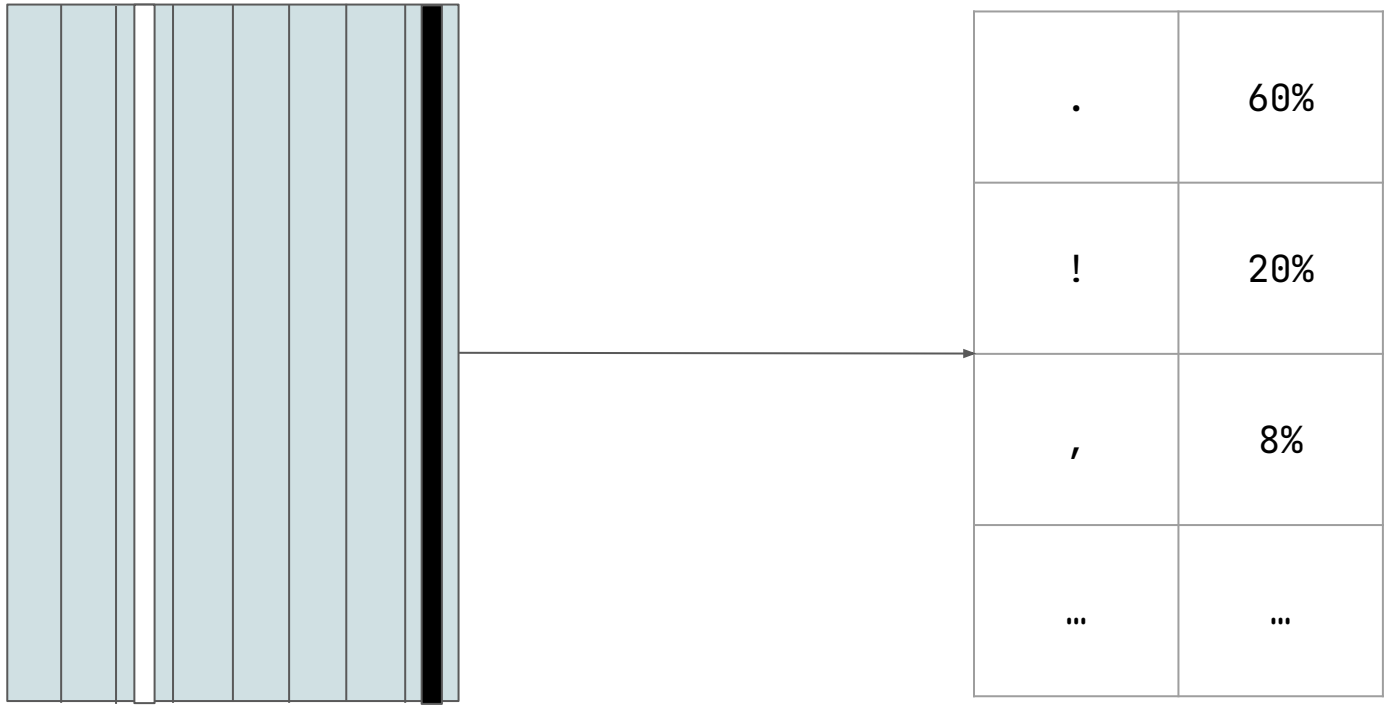


batch tokens dimension
1 4 8

...

Unembed

rld



In conclusion

- Residual stream = shared memory
- Attention: mixes between positions
- MLP: refines per position
- Embed → Transformer Blocks → Unembed = the same skeleton still in use by your favorite LLM

Resources / My path to this understanding

- **A Mathematical Framework for Transformer Circuits:** this paper has a section “Transformer Overview” that has influenced how I think about transformers more than anything else.
- **The papers mentioned in previous slides:** I found working through the GPT-1 paper really valuable. The “*Attention is All You Need*” paper is seminal and many people have created resources centering around it.
- **YouTube:** Serrano.Academy, Neel Nanda, Andrej Karpathy
- **Podcasts:** M*achine Learning Street Talk*, *The Cognitive Revolution*, *Latent Space*, + anything I can find with researchers from the frontier labs
- **Hands on:** I learned a lot from fine-tuning GPT-1
- **Claude and ChatGPT**