Training Transformers

Soumadeep Saha

Language Modelling - Autoregressive

$$[f(w_{T-1}, \ldots, w_1)]_{w_T}$$

$$\approx P(w_T | w_{T-1}, w_{T-2}, \ldots, w_1)$$

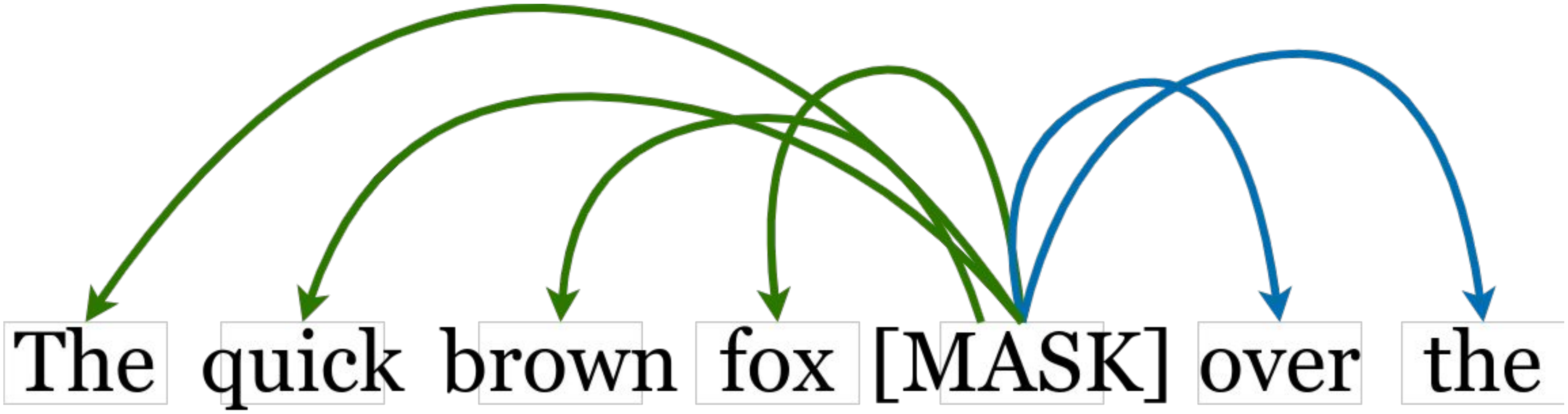
Auto-encoder (Masked Language Modelling)

$$P(w_M | \ldots, w_{M-2}, w_{M-1}, w_{M+1}, \ldots)$$

The quick brown fox jumps over the lazy dog.

The quick brown fox [MASK] over the lazy dog.

Are they the same?



Is it all good?

$$\max_\theta \quad \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)},$$

$$\max_\theta \quad \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)},$$

Is it all good?

$$\max_{\theta} \quad \log p_\theta(\mathbf{x}) = \sum_{t=1}^{T} \log p_\theta(x_t \mid \mathbf{x}_{<t}) = \sum_{t=1}^{T} \log \frac{\exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x_t)\right)}{\sum_{x'} \exp\left(h_\theta(\mathbf{x}_{1:t-1})^\top e(x')\right)},$$

$$\max_{\theta} \quad \log p_\theta(\bar{\mathbf{x}} \mid \hat{\mathbf{x}}) \approx \sum_{t=1}^{T} m_t \log p_\theta(x_t \mid \hat{\mathbf{x}}) = \sum_{t=1}^{T} m_t \log \frac{\exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x_t)\right)}{\sum_{x'} \exp\left(H_\theta(\hat{\mathbf{x}})_t^\top e(x')\right)},$$

$$\hat{\mathbf{x}}) \approx \sum^{T} m$$

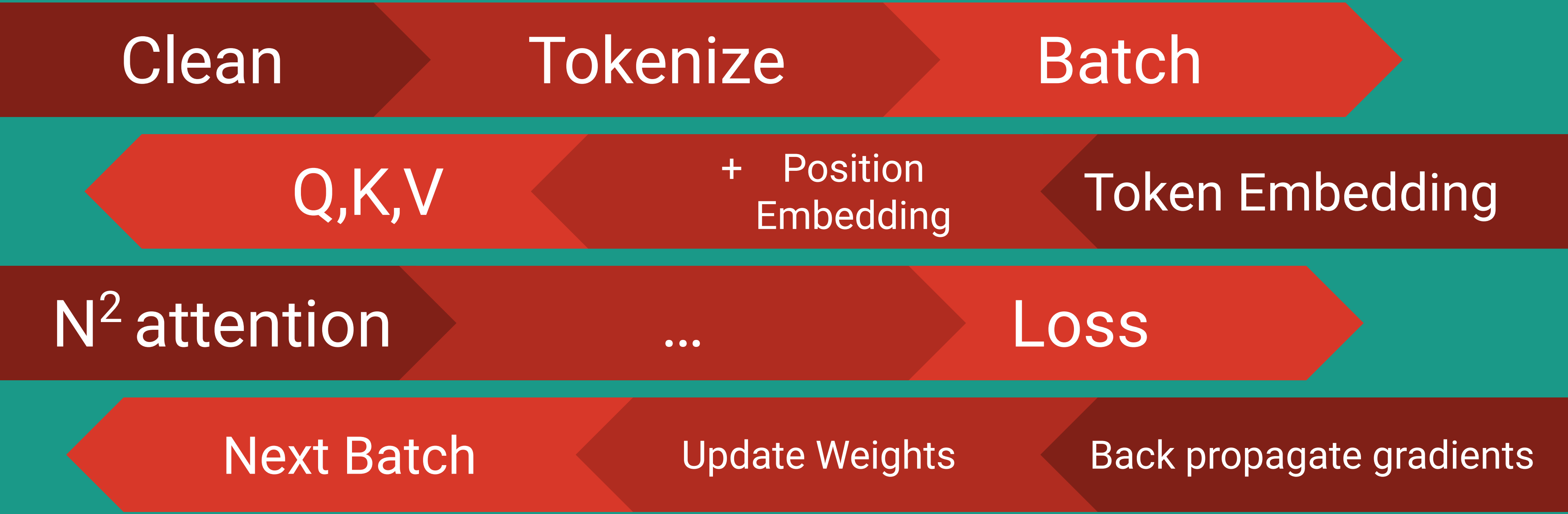
Pros and Cons...

- 15% of [MASK] tokens…
- Independence assumption.

BUT…

- We get bi-directional context

Training pipeline



Walkthrough

The quick brown fox jumps over the lazy dog. 🦊🐶 敏捷的棕色狐狸跳过了懒狗 <a href=”www.google.com”>click

Multilingual model, cased/uncased, etc

the quick brown fox jumps over the lazy dog

Special tokens

[[CLS], the, quick, br#,#own, fox, jumps, over, the, lazy, dog,[SEP]]

Walkthrough

[[CLS], the, quick, br#,#own, fox, jumps, over, the, lazy, dog,[SEP]] **Collect several in a batch [tokenized]**

[[CLS], the, ..., [SEP], [PAD], [PAD], ..., [PAD]]

[[CLS], the, ..., [SEP], [PAD], [PAD], ..., [PAD]]

$$\left[\begin{bmatrix} a_{11} \\ a_{12} \\ \cdots \\ a_{1d} \end{bmatrix}, \begin{bmatrix} a_{21} \\ a_{22} \\ \cdots \\ a_{2d} \end{bmatrix}, \cdots, \begin{bmatrix} a_{k1} \\ a_{k2} \\ \cdots \\ a_{kd} \end{bmatrix} \right]$$

Batch

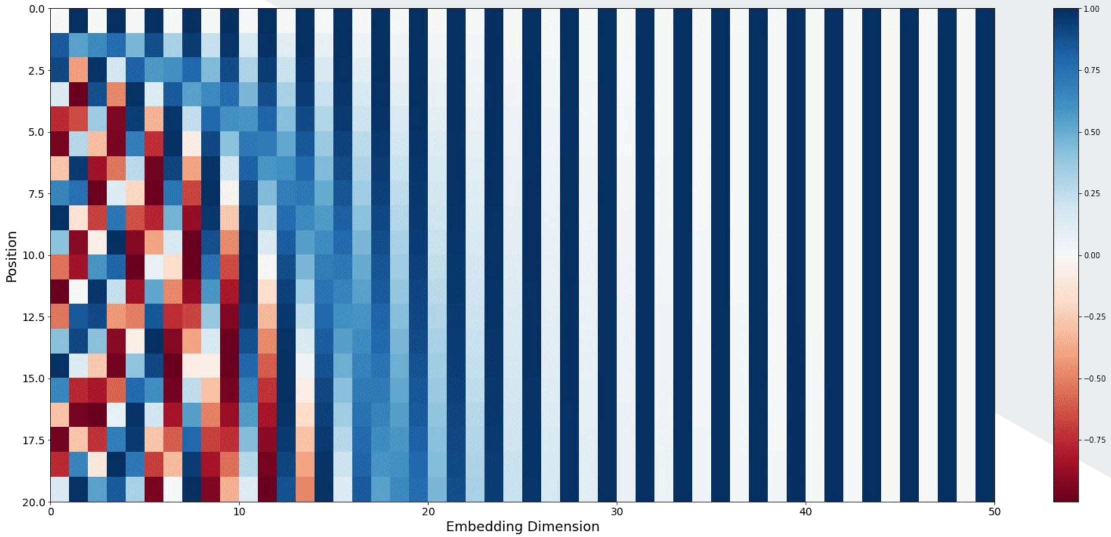
$$\left[\begin{bmatrix} a_{11} \\ a_{12} \\ \cdots \\ a_{1d} \end{bmatrix}, \begin{bmatrix} a_{21} \\ a_{22} \\ \cdots \\ a_{2d} \end{bmatrix}, \dots, \begin{bmatrix} a_{k1} \\ a_{k2} \\ \cdots \\ a_{kd} \end{bmatrix} \right]$$

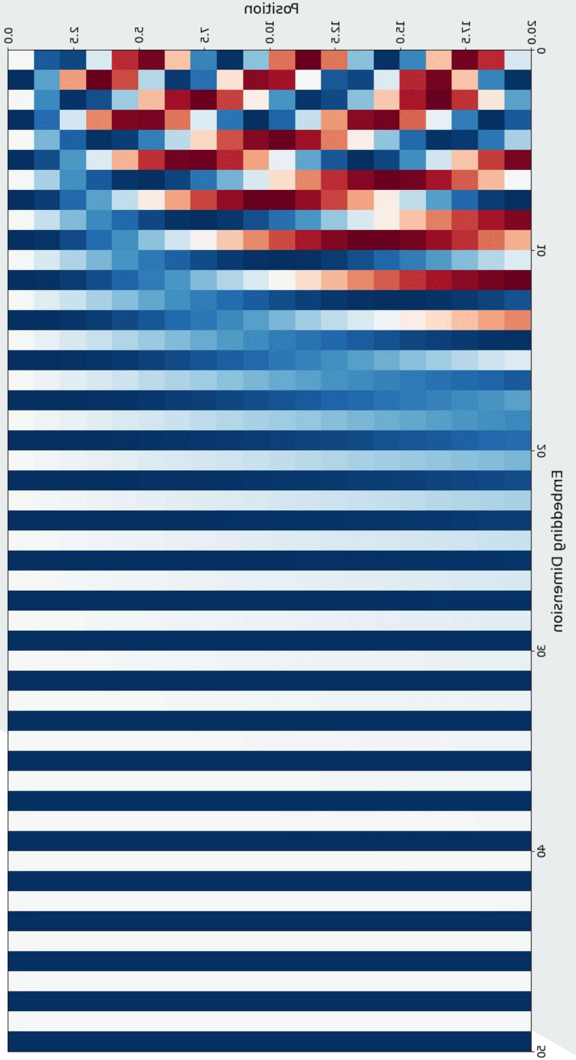
+ Positional embeddings

Positional Embeddings

$$\left[\begin{bmatrix} a_{11} \\ a_{12} \\ \cdots \\ a_{1d} \end{bmatrix}, \begin{bmatrix} a_{21} \\ a_{22} \\ \cdots \\ a_{2d} \end{bmatrix}, \ldots, \begin{bmatrix} a_{k1} \\ a_{k2} \\ \cdots \\ a_{kd} \end{bmatrix} \right]$$

$$\left[\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ldots \right]$$

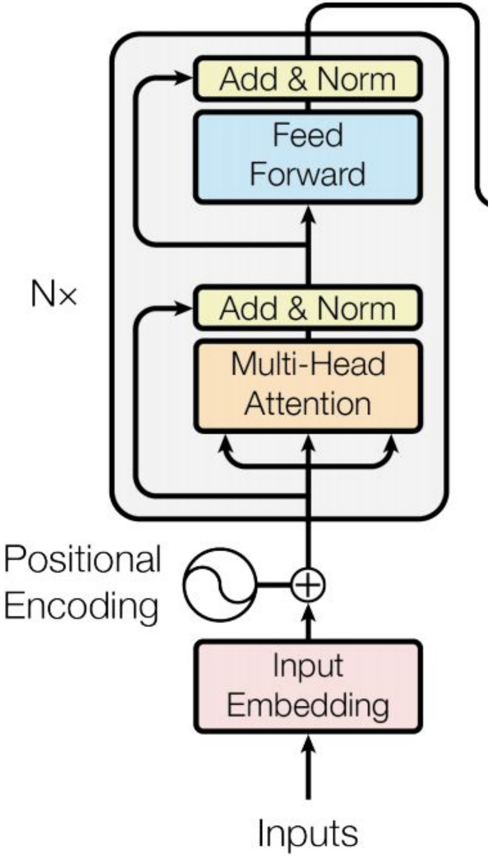




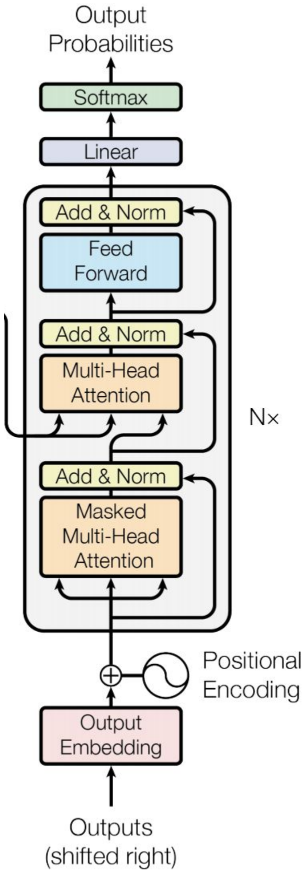
$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

BERT vs GPT

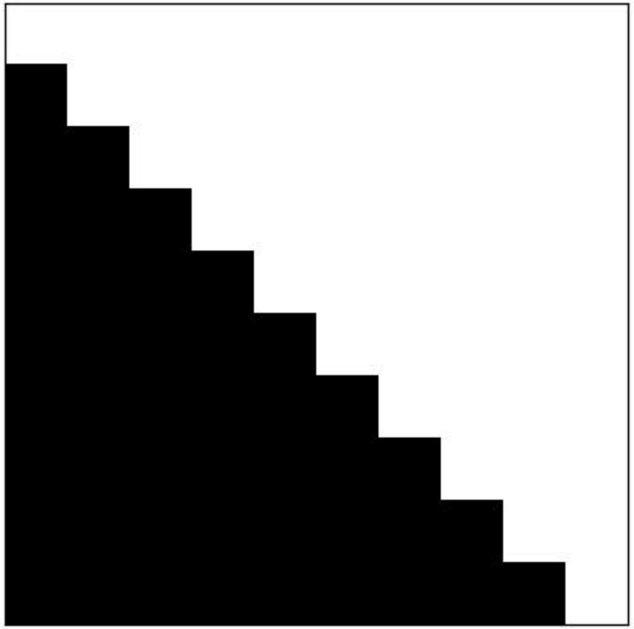


Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Input Embedding
Inputs



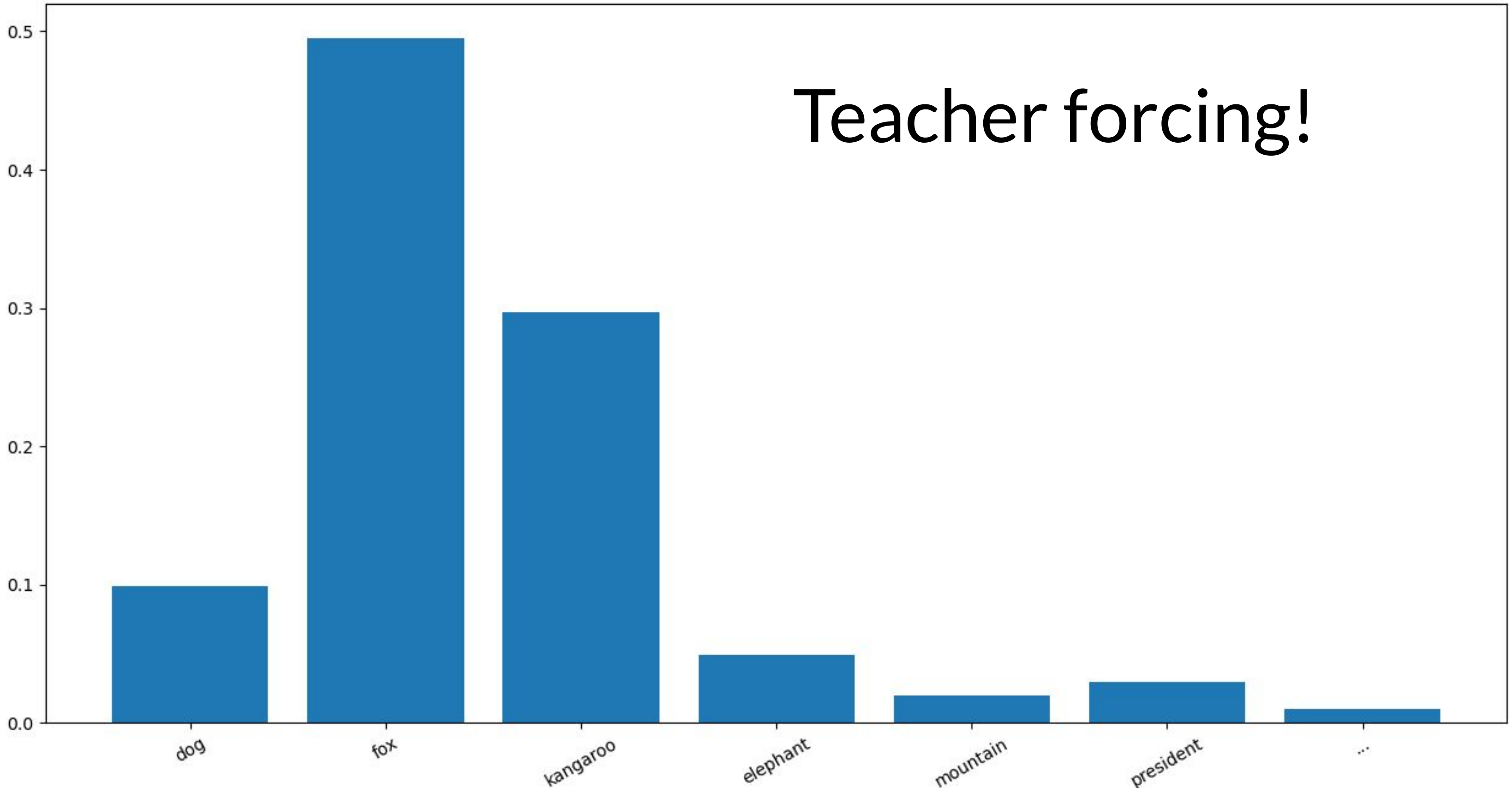
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Positional Encoding
Output Embedding
Outputs (shifted right)

BERT vs GPT



The quick brown fox jumps over the lazy dog.

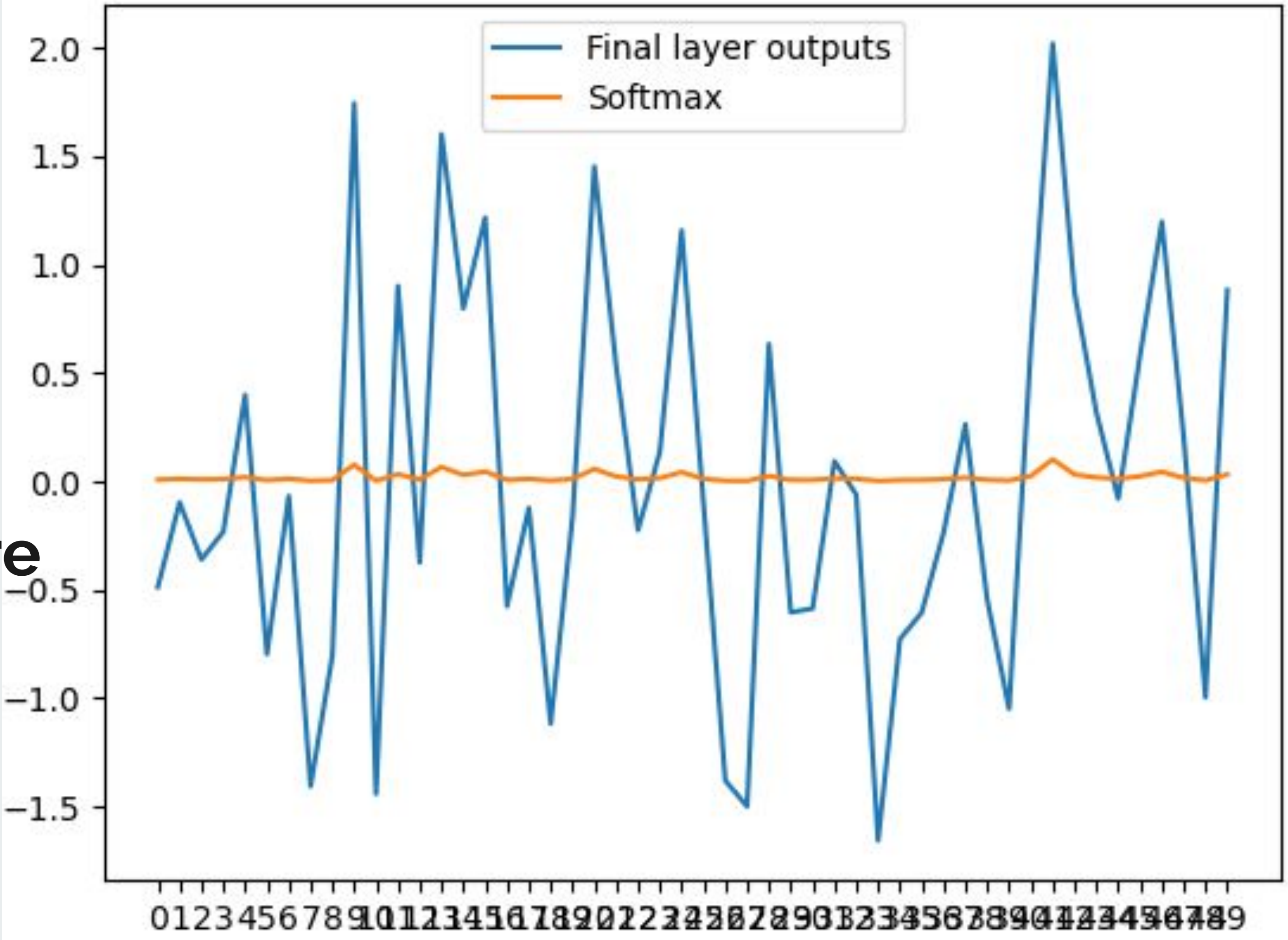
The quick brown fox jumps over the lazy dog.



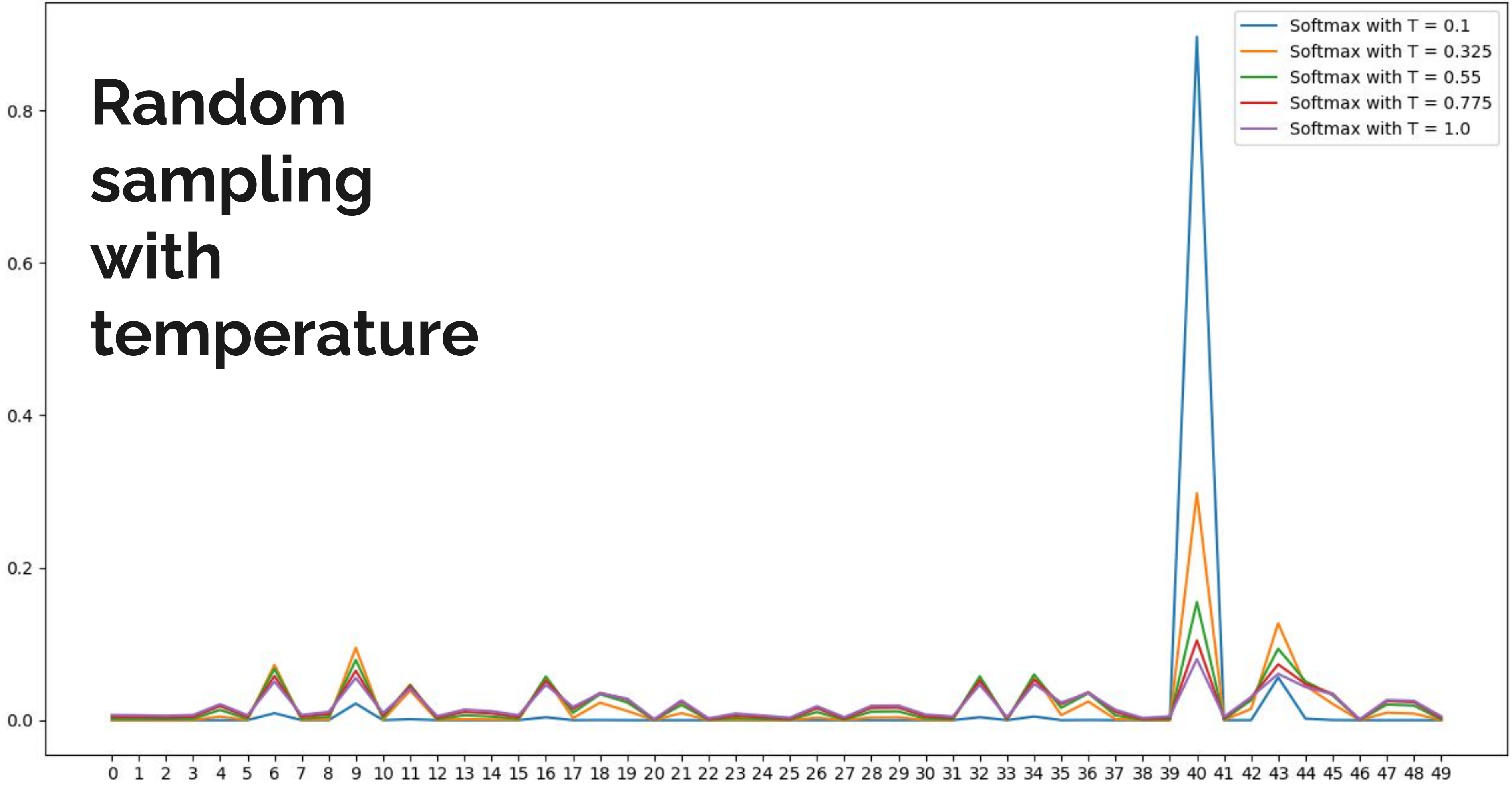
Inference?

- Greedy - pick best answer.
- Random (with Temperature)
- Beam search
- Nucleus Sampling

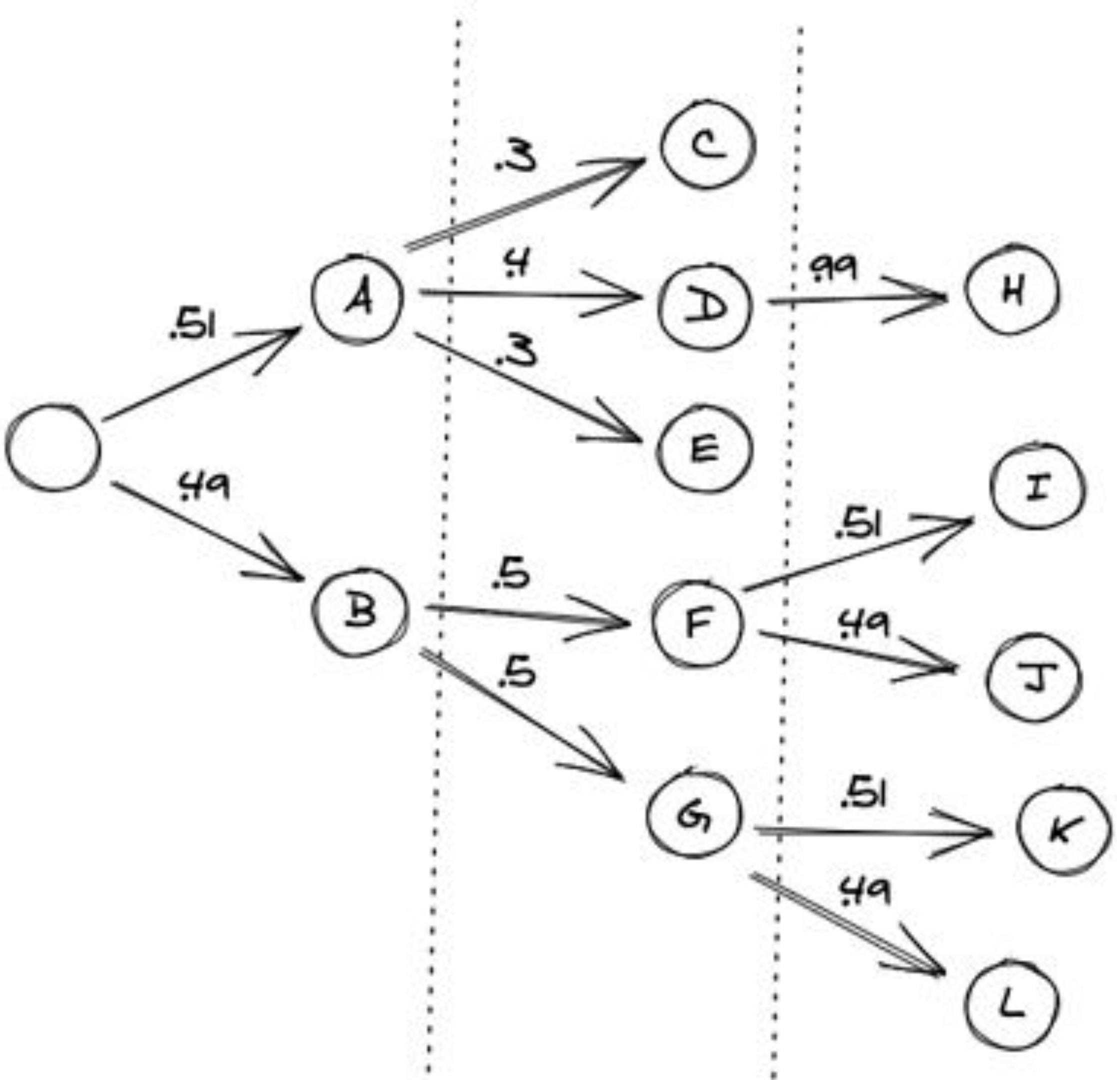
Random sampling with temperature



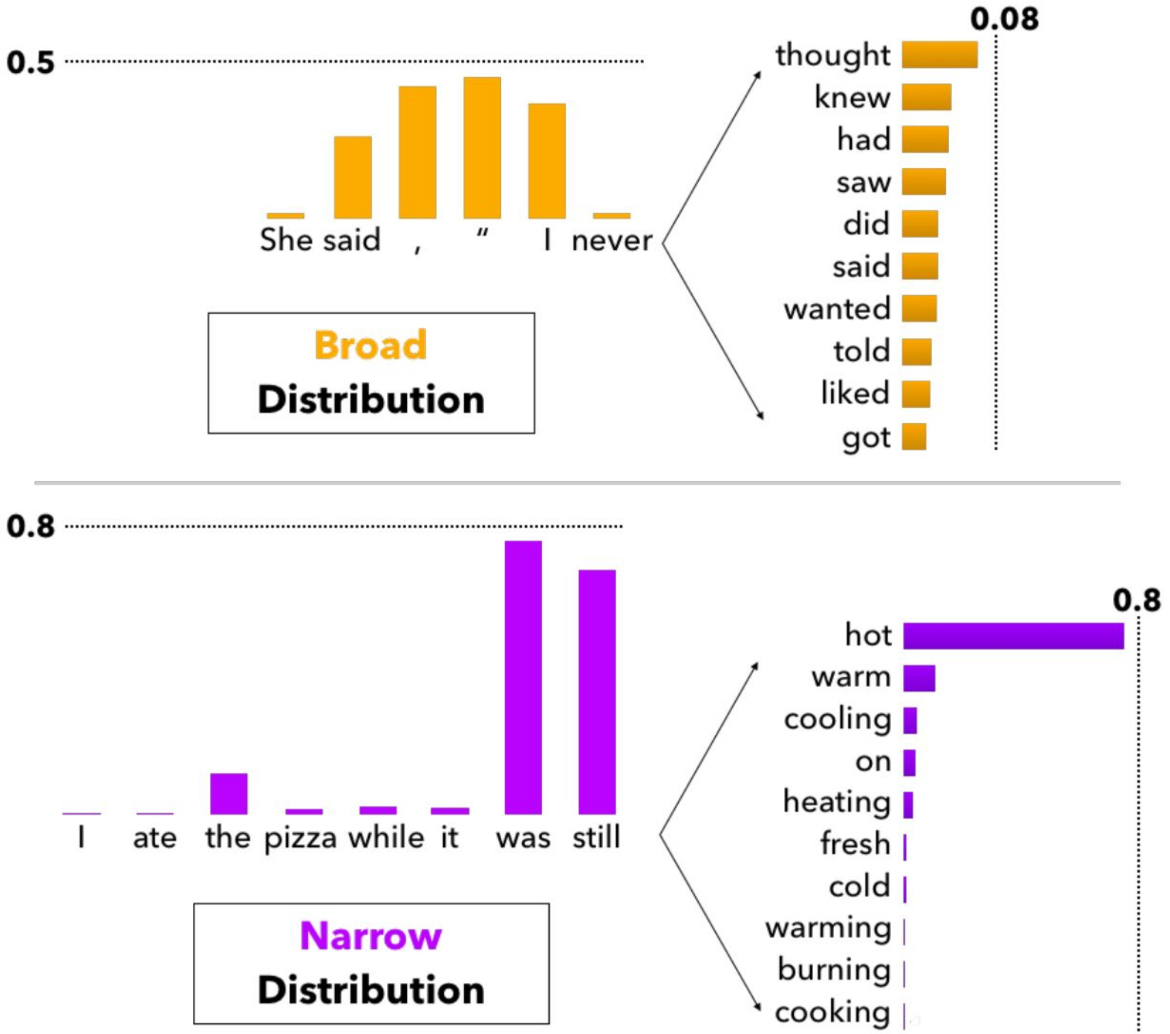
Random sampling with temperature



Beam Search



Nucleus Sampling



Large Foundation Models - Conclusions

- Spoilt for choice [**BERT, LLaMA, Falcon, etc**]
- Size of model - bigger is not always better

  100M - 500B parameters

- Pre-training distribution
- AE vs AR

We'll cover more details in the session on scaling.