Language Models in Cryptanalysis

How do we crack the long ciphers?

Morten Munk

November 2025

AAU CPH - SW9

Contents

1. The papers

1. The papers

X-former Elucidator: Reviving Efficient Attention for Long Context Language Modeling

Miao, et al., 2024

Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention

Katharopoulos, et al., 2020

Long-Short Transformer: Efficient Transformers for Language and Vision

Zhu, et al., 2021

Rethinking Attention with Performers

Choromanski, et al., 2020

Linformer: Self-Attention with Linear Complexity

Wang, et al., 2020

1.1 Why these papers?

Efficient computation during training

- Causal LM inference is fast

Best methods from comparative study

- We don't have time for them all
- Some are more inference focused

Summary

- We care about efficiency during training
- Inference for long cipher struggle due to lack of generalization on long ciphers

2. Recap from last time

Homophonic Substitution Ciphers

- 1:>0 mappings
- English without spaces & punctuation

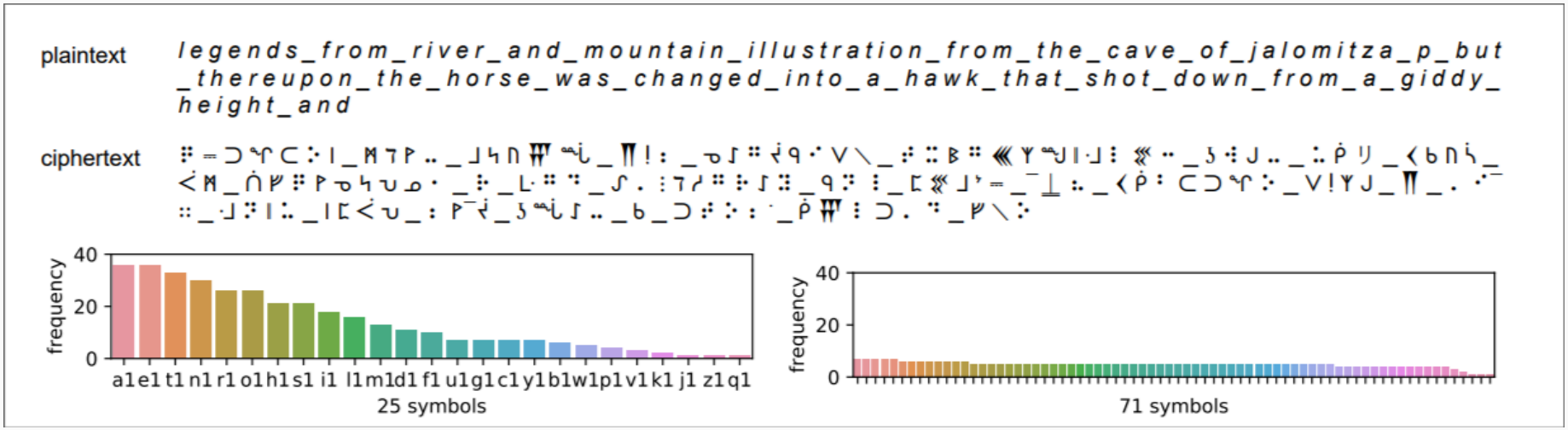

Figure 1: Example of a homophonic substitution cipher. (Kambhatla et al., Findings 2023)

2. Recap from last time

Causal LM

- Learns both cipher & plaintext
- Reads left to right

Seq2seq

- Learns plaintext only
- Bidirectional

Both suffer from $O(N^2)$ attention 😔

#keys	Model	Max Len. 400	700
30-45	Seq-to-Seq	72.30	fail
	PrefixLM	54.73	69.50
	CausalLM (tgt)	29.99	37.20
	CausalLM	**0.40**	**0.21**
40-65	PrefixLM	69.50	54.73
	CausalLM (tgt)	29.99	37.20
	CausalLM	**0.83**	**0.80**
30-85	PrefixLM	70.52	71.82
	CausalLM (tgt)	42.05	42.69
	CausalLM	**2.25**	**2.19**

Figure 2: SER on synthetic HS ciphers. (Kambhatla et al., Findings 2023)

3. Standard Attention Computation

3. Standard Attention Computation

Rows (Queries)
- Token we are looking for

Columns (Keys)
- Token we are looking at

Values (Cells)
- Attention Score

Notice: $(N \times N) = N^2$

Cipher: X Y Z

$$X \to [Y, Z]$$

$$Y \to [X, Z]$$

$$Z \to [X, Y]$$

	X	**Y**	**Z**
X	$X \to X$	$X \to Y$	$X \to Z$
Y	$Y \to X$	$Y \to Y$	$Y \to Z$
Z	$Z \to X$	$Z \to Y$	$Z \to Z$

Table 1: Attention Weight Matrix $(N \times N)$

Why is $O(N^2)$ Ineffecient?

Many attention heads

- Each head has its own matrix

Many forward passes

- Each matrix recomputed at each pass

Causal LM as an example

- 12 layers $\times$ 12 heads $=$ 144 attention heads

Ciphers with 1000s of characters does not scale well 😭

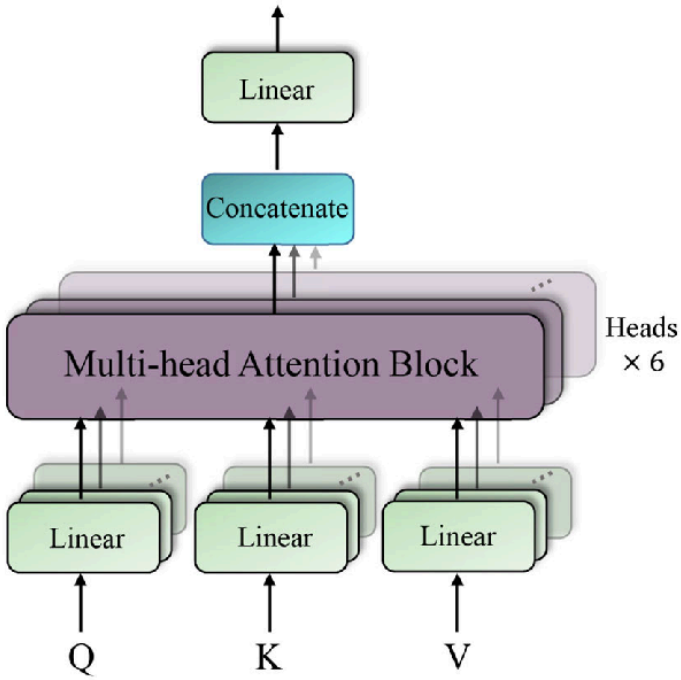

Figure 3: Example of attention with multiple heads. (Yuan et al., 2022)

4. Questions?