Fast Inference from Transformers via Speculative Decoding

Yaniv Leviathan, Matan Kalman, Yossi Matias

ICML 2023

The Gist

- Decode faster from autoregressive models: **2X-3X** in typical scenarios.
- Only different decoding algorithm: **no architecture changes**, **no re-training**.
- **Identical output distribution**.

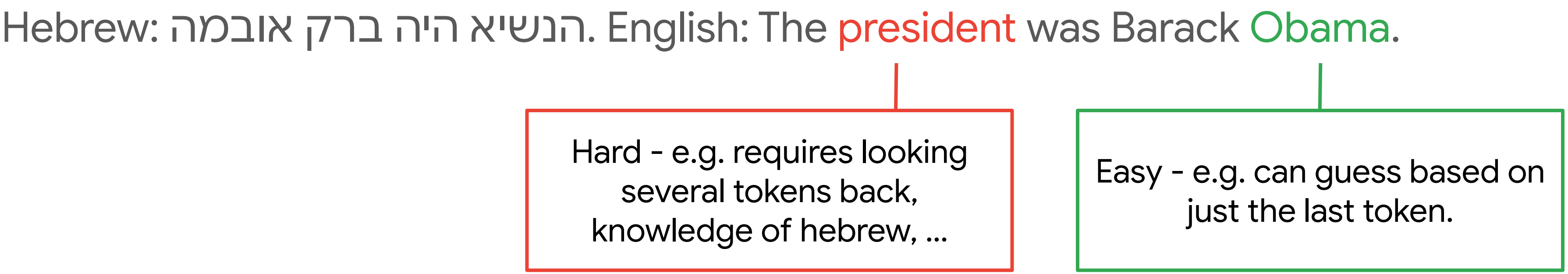
For Autoregressive Models...

Decoding K tokens takes K **serial runs**.

Can we somehow decode several tokens **in parallel**?

Observation 1

Some tokens are easier than others.



Hebrew: הנשיא היה ברק אובמה. English: The president was Barack Obama.

Hard - e.g. requires looking several tokens back, knowledge of hebrew, ...

Easy - e.g. can guess based on just the last token.

Observation 2

Decoding from large Transformers is memory bound.

Hardware can do	Transformers need
XXX Floating point operations per byte read	X Floating point operations per byte read

Contribution 1: Speculative Sampling

Generalization of Speculative Execution to the Stochastic Setting

Speculative Execution

Given **slow** functions f(X) and g(Y):

```
Y = f(X)

Z = g(Y)
```

And given any approximation f*(X) to f(X),

Compute f(X) and g(f*(X)) **in parallel**.

Guarantee identical outputs by rejecting if f(X) ≠ f*(X).

Speculative Sampling

Given **slow** functions f(X) and g(Y):

```
Y ~ f(X)

Z = g(Y)
```

And given any approximation f*(X) to f(X),

Compute f(X) and g(**sample**(f*(X))) **in parallel**.

Guarantee identical distribution by rejecting **w/ some probability** (f(X), f*(X)).

Contribution 2: Speculative Decoding

Application of Speculative Sampling to Decoding from Autoregressive Models

Speculative Decoding

M - auto-regressive model

```
$x_{\leq t} = \mathrm{decode}_M(x_{\leq t-1})$        # f(X)
$x_{\leq t+1} = \mathrm{decode}_M(x_{\leq t})$        # g(Y)
```

Theoretical Highlight 1: Latency Improvement Prediction

The **latency improvement** is a function of:

- How close the approximation model is to the target model (α).
- How fast the approximation model is relative to the target model (c).

Theoretical Highlight 2: Number of Parallel Tokens

We can apply speculative sampling to a **sequence** of slow functions.

We can apply speculative decoding to decode **several** tokens in parallel.

Optimally choosing the number of tokens to attempt to parallelize (γ).

Even off-the-shelf **small models** or **simple heuristics** work well.

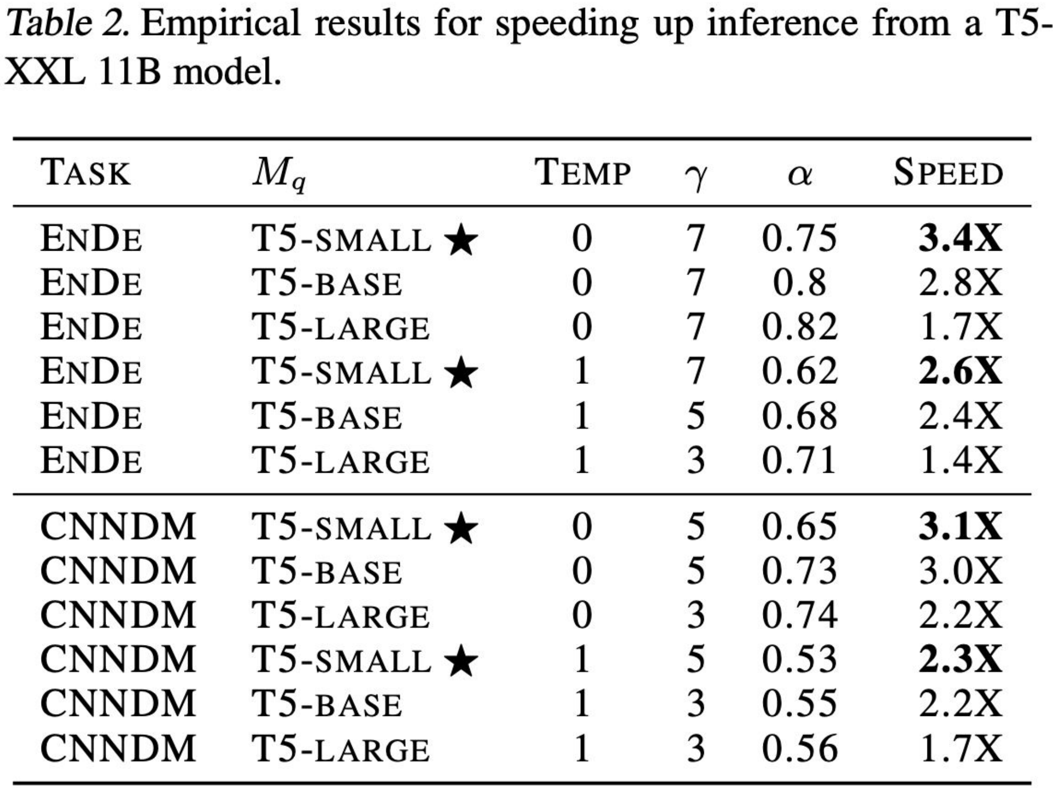


Table 2. Empirical results for speeding up inference from a T5-XXL 11B model.

Task	M_q	Temp	γ	α	Speed
EnDe	T5-small ★	0	7	0.75	**3.4X**
EnDe	T5-base	0	7	0.8	2.8X
EnDe	T5-large	0	7	0.82	1.7X
EnDe	T5-small ★	1	7	0.62	**2.6X**
EnDe	T5-base	1	5	0.68	2.4X
EnDe	T5-large	1	3	0.71	1.4X
CNNDM	T5-small ★	0	5	0.65	**3.1X**
CNNDM	T5-base	0	5	0.73	3.0X
CNNDM	T5-large	0	3	0.74	2.2X
CNNDM	T5-small ★	1	5	0.53	**2.3X**
CNNDM	T5-base	1	3	0.55	2.2X
CNNDM	T5-large	1	3	0.56	1.7X

Thank you!

leviathan@google.com