Promises and Pitfalls of Generative Masked Language Modeling: Theoretical Framework and Practical Guidelines

Yuchen Li[1,2], Alexandre Kirchmeyer[1], Aashay Mehta[1], Yilong Qin[1],
Boris Dadachev[2], Kishore Papineni[2], Sanjiv Kumar[2], Andrej Risteski[1]
([1]CMU [2]Google)

arxiv.org/abs/2407.21046 (ICML 2024)

The autoregressive language model paradigm

Learn an autoregressively parametrized distribution:

$$P_\theta(X_1, X_2, \cdots, X_N) = \prod_{i=1}^{N} P_\theta(X_i \mid X_1, \cdots, X_{i-1})$$

Issues:

1. *Lack of parallelism*

   N sequential steps to generate N tokens

2. *Quality**

- Can't access right-hand context
- No natural way to revise earlier (left) predictions

* Li and Risteski. (ACL 2021)
* Lin et al. (NAACL 2021)
* Bachmann and Nagarajan (arXiv 2024)

Alternative: Generative Masked Language Models*

Non-autoregressive way to generate a sequence*:

- Start w/ pure noise (e.g. masks, random tokens)
- Iteratively refine current guess, s.t. one forward pass updates multiple positions simultaneously.

Bidirectional context. Leverages "parallelism" of transformers for each step.

If # of steps is small, latency is low.

* Jacob Devlin et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
* Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model
* Marjan Ghazvininejad et al. 2019. Mask-predict: Parallel decoding of conditional masked language model
* Jacob Austin. 2021. Structured denoising diffusion models in discrete state-spaces
* Jiatao Gu and Xiang Kong. 2021. Fully non-autoregressive neural machine translation: Tricks of the trade.
* Kartik Goyal et al. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings
* Nikolay Savinov et al. 2022. Step-unrolled denoising autoencoders for text generation

Example of the iterative refinement process

- translate from German to English: Im Fußball geht alles sehr schnell
- human label: Everything moves very fast in football.
- initial decoder hypothesis: <random> <random> <random> …
- decode step 1: Everything football very fast in football.
- decode step 2: Everything is very fast in football.
- decode step 4: Everything is very fast in football.
- decode step 8: Everything is very fast in football.

Example of the iterative refinement process

- human label: Noble Peace Prize winner and former Head of the International Atomic Energy Authority, Mohamed El-Baradei explained that the constitutional draft belongs "on the rubbish tip of history."
- decode step 1: Nobel Peace Prize laureate and ex- of the International Atomic Energy Agency Mohamed ElBaradei said the draft constitution belongson the of rubbish of history".
- decode step 2: Nobel Peace Prize laureate and ex-head of the International Atomic Energy Agency Mohamed El-Baradei said the draft constitution belongs "on the mountain of rubb rub of history".
- decode step 4: Nobel Peace Prize laureate and ex-head of the International Atomic Energy Agency Mohamed El-Baradei said the draft constitution belongs "on the mountain of rubbish in history".
- decode step 8: Nobel Peace Prize laureate and ex-head of the International Atomic Energy Agency Mohamed El-Baradei said the draft constitution belongs "on the mountain of rubbish in history".

Generative Masked Language Models

Training: predict (random) set of tokens, given rest.
In other words, fit $P_\theta(X_S \mid X_{\bar{S}})$

- **Original:** Andrew Carnegie famously said, "My heart is in the work."
- **Masked:** Andrew Carnegie famously [MASK], "My heart is in the [MASK]."

Generation: use the learned conditionals $P_\theta(X_S \mid X_{\bar{S}})$ as input for a Gibbs sampler.

Generative Masked Language Models

Gibbs sampling:

Repeat:

Let current sequence be $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

Pick $\mathrm{S} \subseteq [n]$ uniformly at random.

Sample $\boldsymbol{x}_S{}' \sim P_\theta(\boldsymbol{X}_S = \boldsymbol{x}_S{}' | \boldsymbol{x}_{\bar{S}})$

Update sequence to $\boldsymbol{y} = (\boldsymbol{x}_S{}', \boldsymbol{x}_{\bar{S}})$

This paper

Questions:

How well do we fit *joint* distribution by training to fit the *conditionals*?

Can we use theory to elucidate the design space of losses, training and inference procedures?

Answers:

(1) *A mathematical framework* to analyze training *sample efficiency* & *inference efficiency* of masked language models (MLMs).

(2) *(Not in this talk) Empirical* analysis of critical components & failure modes.*

* Li et al. *Promises and Pitfalls of Generative Masked Language Modeling: Theoretical Framework and Practical Guidelines*. ICML 2024.

Highlights

- “*Dictionary*” between
  - sample complexity of MLM losses (“training efficiency”), and
  - mixing times of Markov Chains (“generation efficiency”)
- Directions towards designing better losses and architectures

Part I: Dictionary b/w sample efficiency and mixing time

Theorem 1 (informal): Sample efficiency of MLM losses can be characterized via mixing time of Gibbs-like sampler.
(E.g., masking random subsets of size k during training
≈ Gibbs sampler that randomizes k coordinates)

Training is sample-efficient when generation is efficient !

Part I: Dictionary b/w sample efficiency and mixing time

Theorem 1 (informal): Sample efficiency of MLM losses can be characterized via mixing time of Gibbs-like sampler.
(E.g., masking random subsets of size k during training
$\approx$ Gibbs sampler that randomizes k coordinates)

Theorem 2 (informal): Masking more is (statistically) better.

Part II: Strong correlations harm sample and inference efficiency

Theorem 3 (informal): Strong dependencies among target positions cause:

(1) Slow generation: slow mixing of Gibbs sampler (*multimodal*)

(2) Slow training: poor sample efficiency (*via Theorem 1*)

(3) A step of Gibbs can't be implemented by parallel decoding Transformers (e.g. a forward pass of BERT*)

Proof idea for (3): Each forward pass of parallel decoding Transformers implements a conditional product distribution

* Jacob Devlin et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Part II: Strong correlations harm sample and inference efficiency

Theorem 3 (informal): Strong dependencies among target positions cause:

(1) Slow generation: slow mixing of Gibbs sampler (*multimodal*)

(2) Slow training: poor sample efficiency (*via Theorem 1*)

(3) A step of Gibbs can't be implemented by parallel decoding Transformers (e.g. a forward pass of BERT*)

Remark 1: Simple toy model to explain "stutter" (common failure mode we observe):

"The dog was walking walking along the road"

Remark 2: Explains why these model work much better for machine translation (generation is "less multimodal", and target-side dependency is weaker)

Future work: ideas to improve losses + samplers

- "Dependent" version of Gibbs sampler where masks are adaptively chosen. (Details in paper)
    - Unclear how to measure "dependence"
    - Preliminary evidence cross-attention is better than self-attention

- Better architectures to implement Markov Chain update in parallel?

* Li et al. *Promises and Pitfalls of Generative Masked Language Modeling: Theoretical Framework and Practical Guidelines*. ICML 2024.