Mask-Predict: Parallel Decoding of Conditional Masked Language Models

Marjan Ghazvininejad*, Omer Levy*, Yinhan Liu*,
Luke Zettlemoyer
Facebook AI Research

EMNLP 2019

Motivation

- Old
  - Most machine translation systems use sequential decoding strategies
  - Words are predicted one by one
  - Autoregressive
- New
  - generates translations in a **constant number** of decoding iterations
  - trained with a masked language model objective
  - Non Autoregressive / parallel decoding

Conditional Masked Language Models

- P($\mathrm{Y_{mask}} \mid \mathrm{X}, \mathrm{Y_{obs}}$)
- Assumption: tokens $\mathrm{Y_{mask}}$ are conditionally independent
- Model is implicitly conditioned on length of target sequence
- N = | $\mathrm{Y_{mask}}$ | + | $\mathrm{Y_{obs}}$ |

Architecture

- Transformer
- no self-attention mask that prevents left-to-right decoders from attending on future tokens
- decoder is bi-directional

Training Objective

- Sample number of masked tokens
- Replace inputs of tokens $\mathbf{Y_{mask}}$ with a special MASK token
- Optimize CMLM for cross-entropy loss over every token in $\mathbf{Y_{mask}}$

Predicting Target Sequence Length

- Traditionally: predict EOS (end of sentence) token
- CMLMs: must know length in advance
- special LENGTH token is added to encoder (akin to CLS in BERT)

Decoding with Mask-Predict

- t = 0
  - mask all tokens
- t > 0
  - Mask: mask n tokens with lowest probability scores:

$$Y_{mask}^{(t)} = \arg\min_{i}(p_i, n)$$

$$Y_{obs}^{(t)} = Y \setminus Y_{mask}^{(t)}$$

$$n = N \cdot \frac{T-t}{T}$$

  - Predict:

$$y_i^{(t)} = \arg\max_{w} P(y_i = w | X, Y_{obs}^{(t)})$$

$$p_i^{(t)} = \max_{w} P(y_i = w | X, Y_{obs}^{(t)})$$

Decoding with Mask-Predict

src	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed on November 20th .

Figure 1: An example from the WMT'14 DE-EN validation set that illustrates how mask-predict generates text. At each iteration, the highlighted tokens are masked and repredicted, conditioned on the other tokens in the sequence.

Deciding Target Sequence Length

- First: compute CMLM's encoder
- Then: use LENGTH token's encoding to predict distribution over sequence's length
- Select top ℓ length candidates with highest probabilities
- Decode same example with different lengths in parallel
- Select the sequence with highest average log-probability:

$$\frac{1}{N}\sum \log p_i^{(T)}$$

Experiments

- WMT'14 EN-DE, WMT'16 EN-RO, WMT'17 EN-ZH
- Mostly standard parameters for transformers for baseline
- Weight initialization according to BERT
- **Knowledge distillation**
  - Train CMLMs on translations produced by a standard left-to-right transformer model

Model	Dimensions (Model/Hidden)	Iterations	WMT'14 EN-DE	WMT'14 DE-EN	WMT'16 EN-RO	WMT'16 RO-EN
NAT w/ Fertility (Gu et al., 2018)	512/512	1	19.17	23.20	29.79	31.44
CTC Loss (Libovický and Helcl, 2018)	512/4096	1	17.68	19.80	19.93	24.71
Iterative Refinement (Lee et al., 2018)	512/512	1	13.91	16.77	24.45	25.73
	512/512	10	21.61	25.48	29.32	30.19
(Dynamic #Iterations)	512/512	?	21.54	25.43	29.66	30.30
Small CMLM with Mask-Predict	512/512	1	15.06	19.26	20.12	20.36
	512/512	4	**24.17**	**28.55**	**30.00**	30.43
	512/512	10	**25.51**	**29.47**	**31.65**	**32.27**
Base CMLM with Mask-Predict	512/2048	1	18.05	21.83	27.32	28.20
	512/2048	4	**25.94**	**29.90**	**32.53**	**33.23**
	512/2048	10	**27.03**	**30.53**	**33.08**	**33.31**
Base Transformer (Vaswani et al., 2017)	512/2048	N	27.30	— —	— —	— —
Base Transformer (Our Implementation)	512/2048	N	27.74	31.09	34.28	33.99
Base Transformer (+Distillation)	512/2048	N	27.86	31.07	— —	— —
Large Transformer (Vaswani et al., 2017)	1024/4096	N	28.40	— —	— —	— —
Large Transformer (Our Implementation)	1024/4096	N	28.60	31.71	— —	— —

Table 1: The performance (BLEU) of CMLMs with mask-predict, compared to other parallel decoding machine translation methods. The standard (sequential) transformer is shown for reference. Bold numbers indicate state-of-the-art performance among parallel decoding methods.

Model	Dimensions (Model/Hidden)	Iterations	WMT'17 EN-ZH	WMT'17 ZH-EN
Base CMLM with Mask-Predict	512/2048	1	24.23	13.64
	512/2048	4	32.63	21.90
	512/2048	10	33.19	23.21
Base Transformer (Our Implementation)	512/2048	N	34.31	23.74
Base Transformer (+Distillation)	512/2048	N	34.44	23.99
Large Transformer (Our Implementation)	1024/4096	N	35.01	24.65

Table 2: The performance (BLEU) of CMLMs with mask-predict, compared to the standard (sequential) transformer on WMT' 17 EN-ZH.

Decoding Speed

- Base transformer for baseline system with beam search (EN-DE)
- Also use greedy search for faster but less accurate baseline
- Varied number of mask-predict iterations (T = 4; :::10)
- Varied number of length candidates (ℓ = 1; 2; 3)
- Measure performance (BLEU) and wall time (seconds)
- Calculate relative decoding speed-up (CMLM time / baseline time)

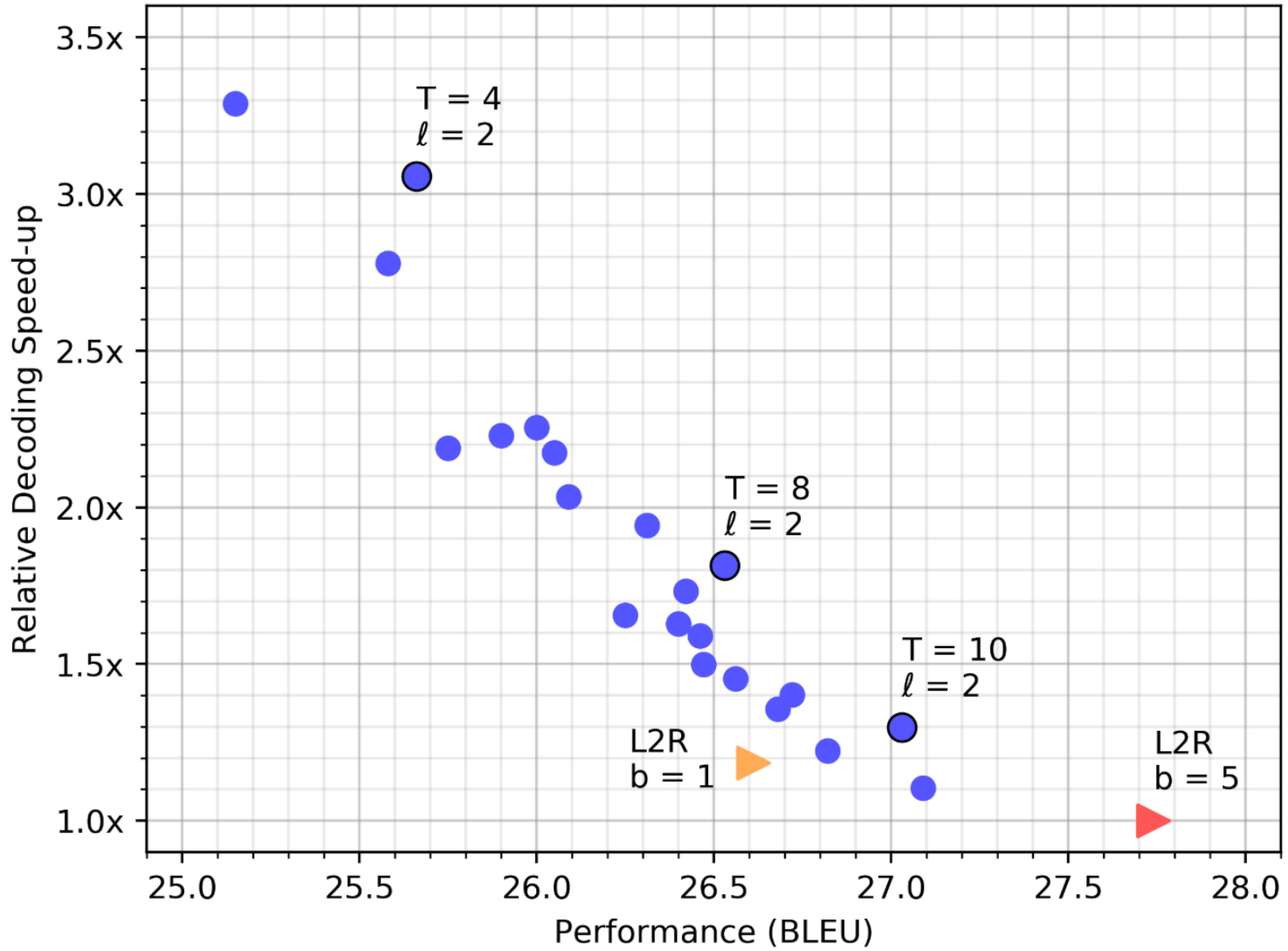

Figure 2: The trade-off between speed-up and translation quality of a base CMLM with mask-predict, compared to the standard sequentially-decoded base transformer on the WMT'14 EN-DE test set, with beam sizes $b = 1$ (orange triangle) and $b = 5$ (red triangle). Each blue circle represents a mask-predict decoding run with a different number of iterations ($T = 4, \ldots, 10$) and length candidates ($\ell = 1, 2, 3$).

Qualitative Analysis

- Why are multiple iterations necessary?
- Do longer sequences need more iterations?
- Do more length candidates help?
- Is model distillation necessary?

Why are multiple iterations necessary?

Iterations	WMT'14 EN-DE		WMT'16 EN-RO	
	BLEU	Reps	BLEU	Reps
$T = 1$	18.05	16.72%	27.32	9.34%
$T = 2$	22.91	5.40%	31.08	2.82%
$T = 3$	24.99	2.03%	32.19	1.26%
$T = 4$	25.94	1.07%	32.53	0.87%
$T = 5$	26.30	0.72%	32.62	0.61%

Table 3: The performance (BLEU) and percentage of repeating tokens when decoding with a different number of mask-predict iterations (T).

- multi-modality problem (token repetitions)

Do longer sequences need more iterations?

	$T = 4$	$T = 10$	$T = N$
$1 \leq N < 10$	21.8	22.4	22.4
$10 \leq N < 20$	24.6	25.9	26.0
$20 \leq N < 30$	24.9	26.7	27.1
$30 \leq N < 40$	24.9	26.7	27.6
$40 \leq N$	25.0	27.5	28.1

Table 4: The performance (BLEU) of base CMLM with different amounts of mask-predict iterations (T) on WMT'14 EN-DE, bucketed by target sequence length (N). Decoding with $\ell = 1$ length candidates.

Do more length candidates help?

Length Candidates	WMT'14 EN-DE BLEU	WMT'14 EN-DE LP	WMT'16 EN-RO BLEU	WMT'16 EN-RO LP
$\ell = 1$	26.56	16.1%	32.75	13.8%
$\ell = 2$	27.03	30.6%	33.06	26.1%
$\ell = 3$	**27.09**	43.1%	**33.11**	39.6%
$\ell = 4$	**27.09**	53.1%	32.13	49.2%
$\ell = 5$	27.03	62.2%	33.08	57.5%
$\ell = 6$	26.91	69.5%	32.91	64.3%
$\ell = 7$	26.71	75.5%	32.75	70.4%
$\ell = 8$	26.59	80.3%	32.50	74.6%
$\ell = 9$	26.42	83.8%	32.09	78.3%
Gold	27.27	—	33.20	—

Table 5: The performance (BLEU) of base CMLM with 10 mask-predict iterations ($T = 10$), varied by the number of length candidates (ℓ), compared to decoding with the reference target length (Gold). Length precision (LP) is the percentage of examples that contain the correct length as one of their candidates.

Is model distillation necessary?

Iterations	WMT'14 EN-DE		WMT'16 EN-RO	
	Raw	Dist	Raw	Dist
$T = 1$	10.64	**18.05**	21.22	**27.32**
$T = 4$	22.25	**25.94**	31.40	**32.53**
$T = 10$	24.61	**27.03**	32.86	**33.08**

Table 6: The performance (BLEU) of base CMLM, trained with either raw data (Raw) or knowledge distillation from an autoregressive model (Dist).

Conclusion

- Approach outperforms previous parallel decoding methods
- Approaches the performance of sequential autoregressive models (decoding faster)
- Problem: need to condition on the target's length
- Problem: dependence on knowledge distillation