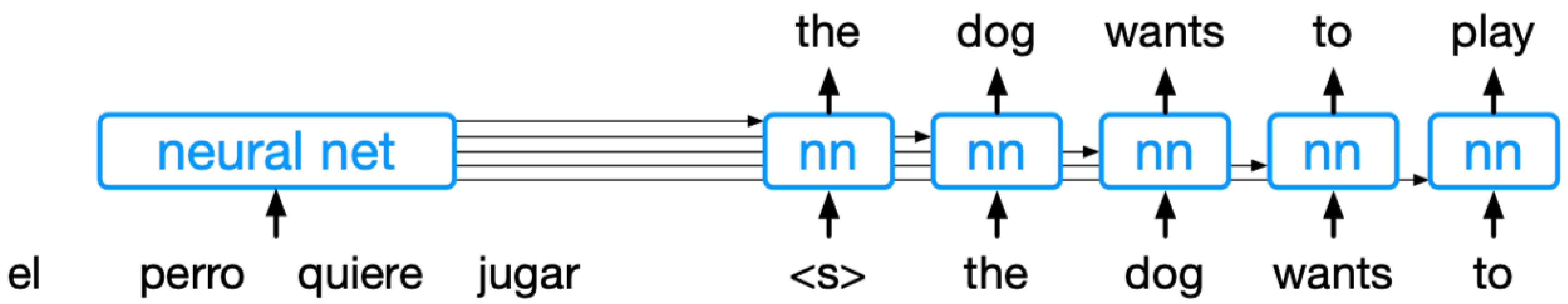
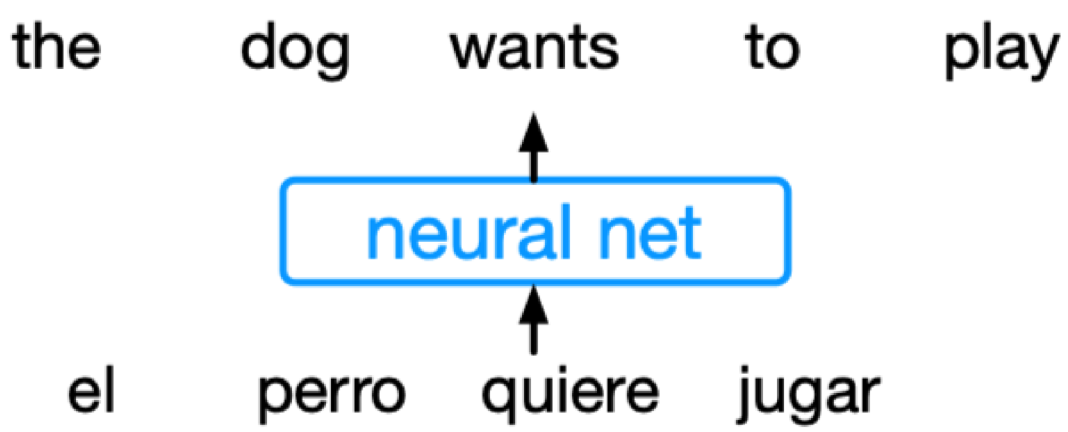
Mask-Predict: Parallel Decoding of Conditional Masked Language Models

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

4 September 2019

Machine translation





autoregressive

non-autoregressive

O(n)

O(1)

Multimodality Problem

Thank you very much

Vielen Dank

Danke schön

CMLM - Conditional Masked Language Models:

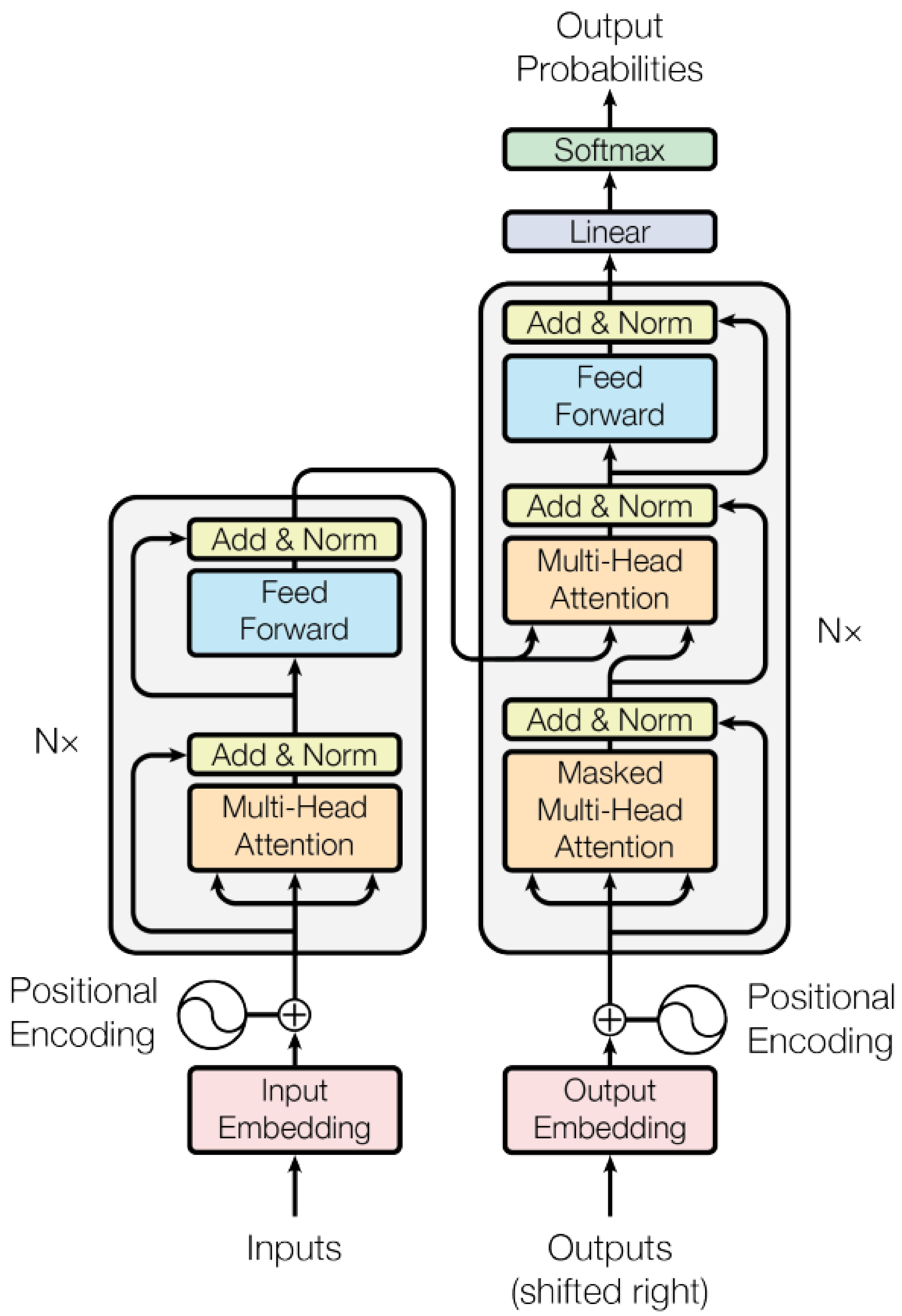


Figure 1: The Transformer - model architecture.

- Encoder-decoder transformer для машинного перевода
- Удалили self-attention слои в декодере, которые не позволяют смотреть на токены справа
- Получили bi-directional декодер

CMLM - Conditional Masked Language Models:

- Маскируем k токенов target sequence = Y_mask
- Основывается на предположении, что Y_mask не зависят друг от друга при условии X, Y_obs
- Предсказываем замаскированные таргеты, смотря на X, Y_obs

Обучение

- маскируем k токенов $k \sim \text{Uniform}(1, N)$
- предсказываем их параллельно, учитывая весь контекст $P(y|X, Y\_obs)$
- предсказываем target length в энкодере

Decoding algorithm: Parallel Decoding with Mask-Predict

Инициализация:

- получить LEN токен
- заменить все токены в таргете на маски
- делаем предсказание с argmax параллельно

Mask-Predict итерация:

- маскируем токены, в которых меньше уверены
- предсказываем параллельно с argmax

Сколько итераций?

- Constant (1-10)
- Function log(n), sqrt(n), n

Сколько слов маскировать?

- $n = N * (T - t) / T$

Пример работы модели

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 .

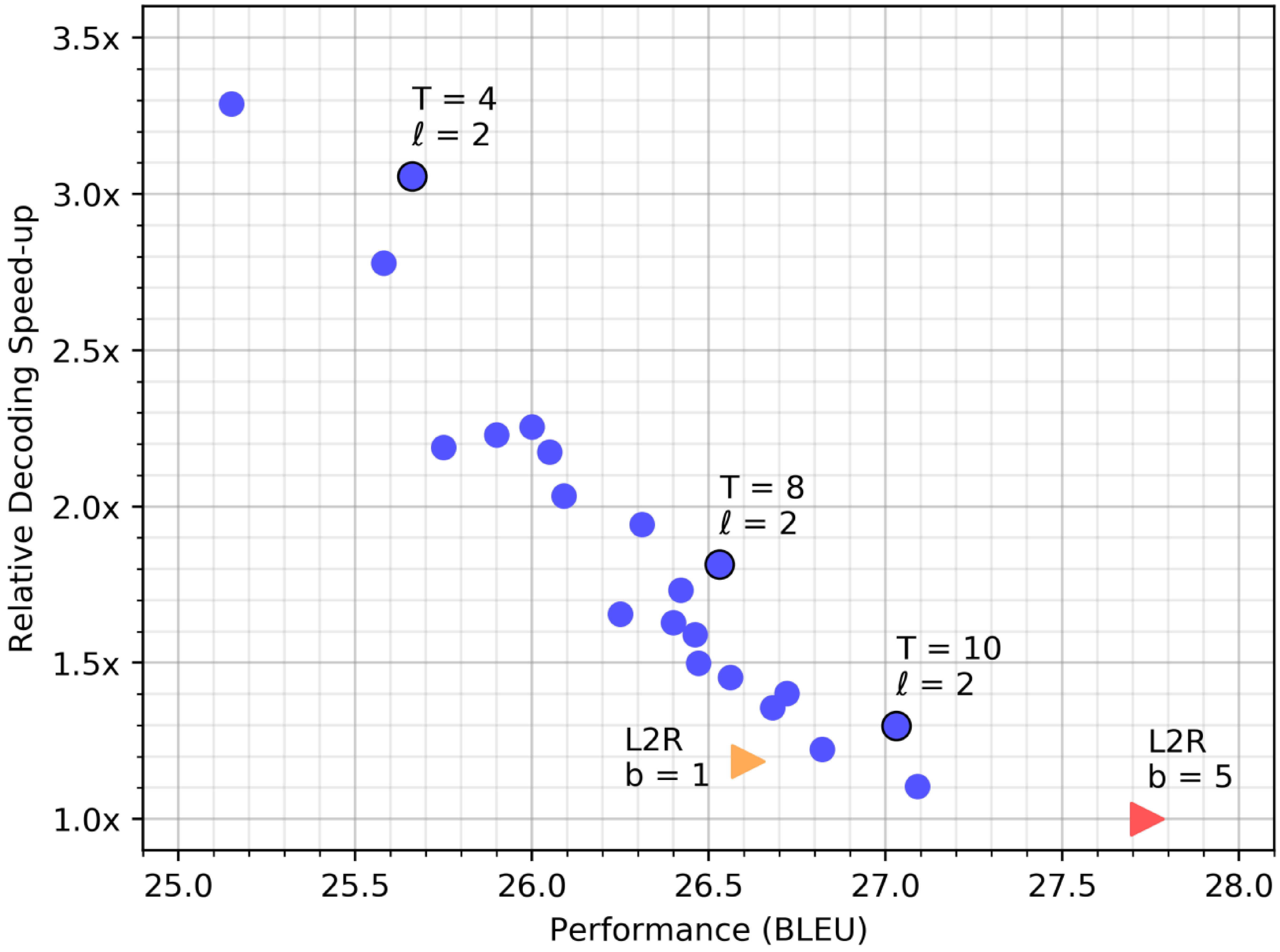
Figure 1: An example from the WMT'14 DE-EN validation set that illustrates how mask-predict generates text. At each iteration, the highlighted tokens are masked and repredicted, conditioned on the other tokens in the sequence.

Качество перевода

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.

Decoding Speed



Why Are Multiple Iterations Necessary?

- снижает повторы токенов в генерации - multimodality problem

Iterations	WMT'14 EN-DE BLEU	WMT'14 EN-DE Reps	WMT'16 EN-RO BLEU	WMT'16 EN-RO 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).

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).