

# Parallel Decoding of Conditional Masked Language Models

Дарья Виноградова

09.03.2021

# Мотивация

- обычно генерируем перевод слева направо авторегрессионно
- замаскируем некоторые слова таргетного перевода и научим модель их предсказывать, основываясь на входном тексте и куске перевода
- можно сгенерировать перевод для всех слов независимо, а потом повторять процедуры для подмножества слов, в котором модель не уверена
- немного хуже работает, зато можно параллелить вычисления
- предполагаем, что предсказываемые слова зависят лишь от входного текста и куска предсказанного (независимы между собой)

# Модель

- CMLM - модель, предсказывающая замаскированное подмножество слов
- в основе - трансформер
- отличие от классического трансформера: смотрим в обе стороны, двигаясь по декодированной последовательности
- дополнительно предсказываем длину перевода (по выходу энкодера)

# Обучение

- маскируем случайное количество слов таргета, предсказываем их
- считаем лосс только для слов маски

# Применение

- предсказываем длину
- предсказываем перевод такой длины параллельно
- фиксированное число итераций перепредсказываем подмножество слов, в которых модель не уверена

# Пример

---

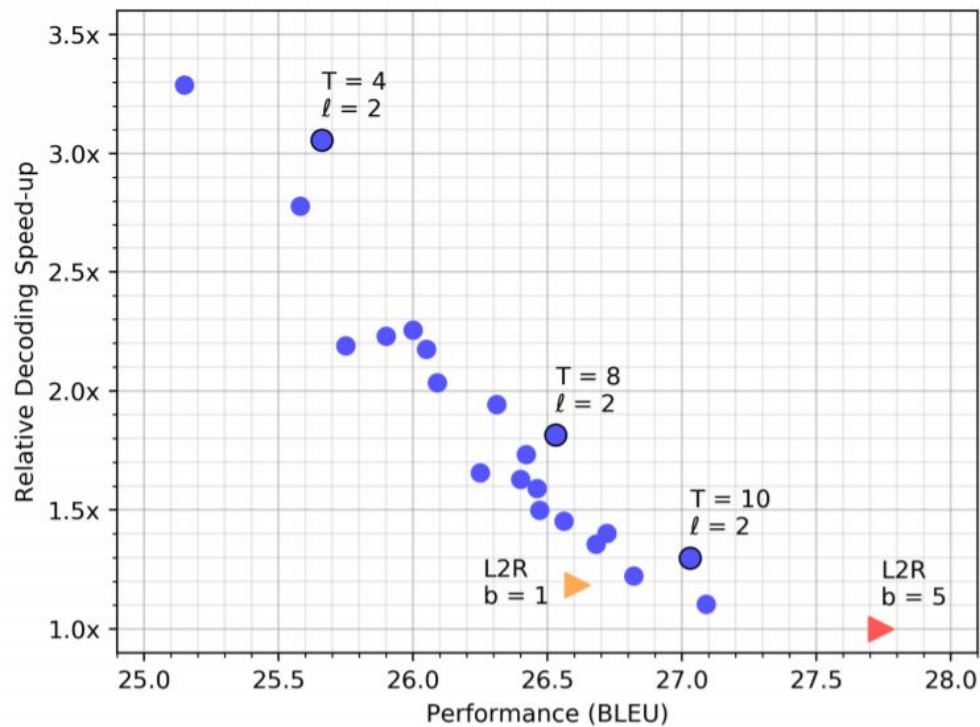
<i>src</i>	Der Abzug der franzsischen Kampftruppen wurde am 20. November abgeschlossen .
$t = 0$	The departure of the French combat completed completed on 20 November .
$t = 1$	The departure of French combat troops was completed on 20 November .
$t = 2$	The withdrawal of French combat troops was completed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

---

# Эксперименты

Model	Dimensions (Model/Hidden)	Iterations	WMT'14		WMT'16	
			EN-DE	DE-EN	EN-RO	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
<hr/>						
<i>Small CMLM with Mask-Predict</i>	512/512	1	15.06	19.26	20.12	20.36
	512/512	4	<b>24.17</b>	<b>28.55</b>	<b>30.00</b>	30.43
	512/512	10	<b>25.51</b>	<b>29.47</b>	<b>31.65</b>	<b>32.27</b>
<i>Base CMLM with Mask-Predict</i>	512/2048	1	18.05	21.83	27.32	28.20
	512/2048	4	<b>25.94</b>	<b>29.90</b>	<b>32.53</b>	<b>33.23</b>
	512/2048	10	<b>27.03</b>	<b>30.53</b>	<b>33.08</b>	<b>33.31</b>
<hr/>						
Base Transformer (Vaswani et al., 2017)	512/2048	<i>N</i>	27.30	— —	— —	— —
Base Transformer (Our Implementation)	512/2048	<i>N</i>	27.74	31.09	34.28	33.99
Base Transformer (+Distillation)	512/2048	<i>N</i>	27.86	31.07	— —	— —
Large Transformer (Vaswani et al., 2017)	1024/4096	<i>N</i>	28.40	— —	— —	— —
Large Transformer (Our Implementation)	1024/4096	<i>N</i>	28.60	31.71	— —	— —

# Сравнение скорости работы



# Помогает ли увеличение числа итераций?

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%



Нужно ли больше итераций для длинных последовательностей?

	$T = 4$	$T = 10$	$T = N$
$1 \leq N < 10$	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

# Как влияет количество длин-кандидатов?

Length Candidates	WMT'14 EN-DE		WMT'16 EN-RO	
	BLEU	LP	BLEU	LP
$\ell = 1$	26.56	16.1%	32.75	13.8%
$\ell = 2$	27.03	30.6%	33.06	26.1%
$\ell = 3$	<b>27.09</b>	43.1%	<b>33.11</b>	39.6%
$\ell = 4$	<b>27.09</b>	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	—

# Как влияет количество длин-кандидатов?

Length Candidates	WMT'14 EN-DE		WMT'16 EN-RO	
	BLEU	LP	BLEU	LP
$\ell = 1$	26.56	16.1%	32.75	13.8%
$\ell = 2$	27.03	30.6%	33.06	26.1%
$\ell = 3$	<b>27.09</b>	43.1%	<b>33.11</b>	39.6%
$\ell = 4$	<b>27.09</b>	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	—

# Есть ли смысл в дистилляции?

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

# Источники

<https://arxiv.org/pdf/1904.09324.pdf> - Mask-Predict: Parallel Decoding of Conditional Masked Language Models