Mask-Predict: Parallel Decoding of Conditional Masked Language Models

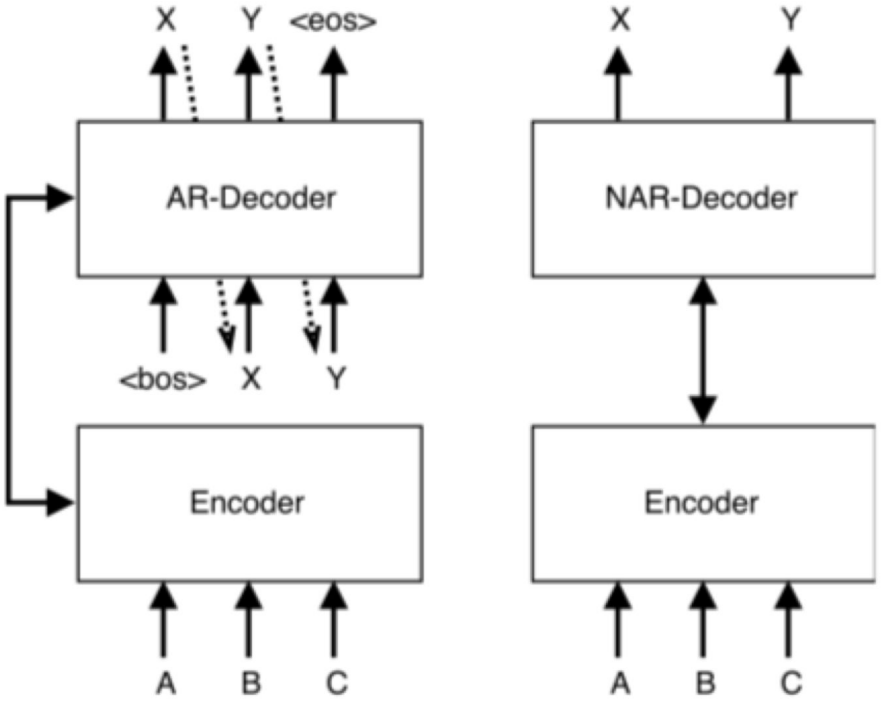
Marjan Ghazvininejad Omer Levy Yinhan Liu Luke Zettlemoyer

Козлов Василий 171

Autoregressive vs non-autoregressive

Авторегрессивный подход
O(n)

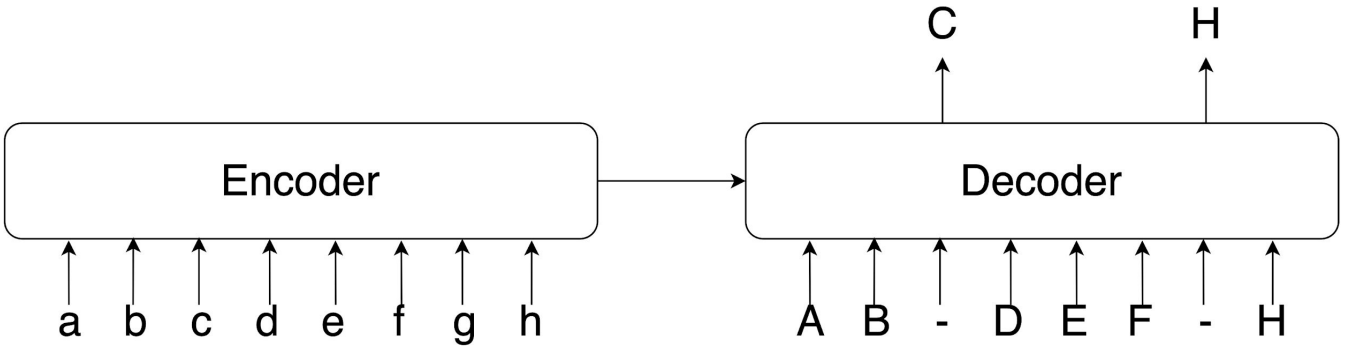
Неавторегрессивный подход
O(1)



Conditional Masked Language Model

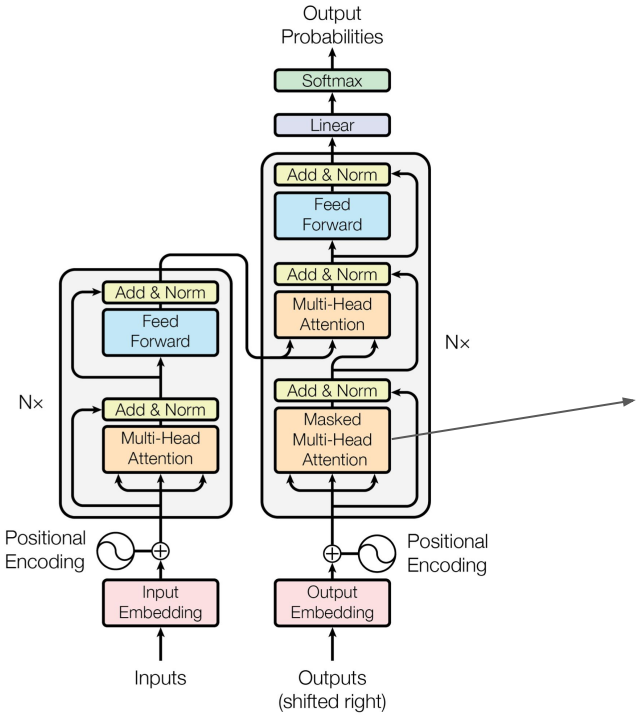
Определение: предсказывает набор Y_{mask} , зная исходный текст X и Y_{obs}.

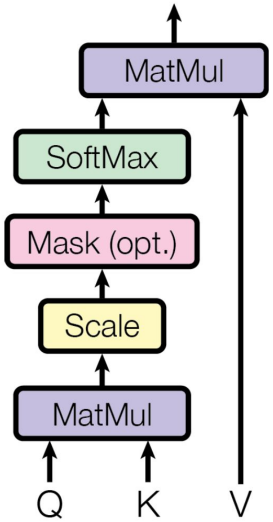
- Полученные токены независимы
- Предсказывает вероятности $P(y|X, Y_{obs})$ для каждого y из Y_{mask}
- Модель должна знать длину $N = |Y_{obs}| + |Y_{mask}|$



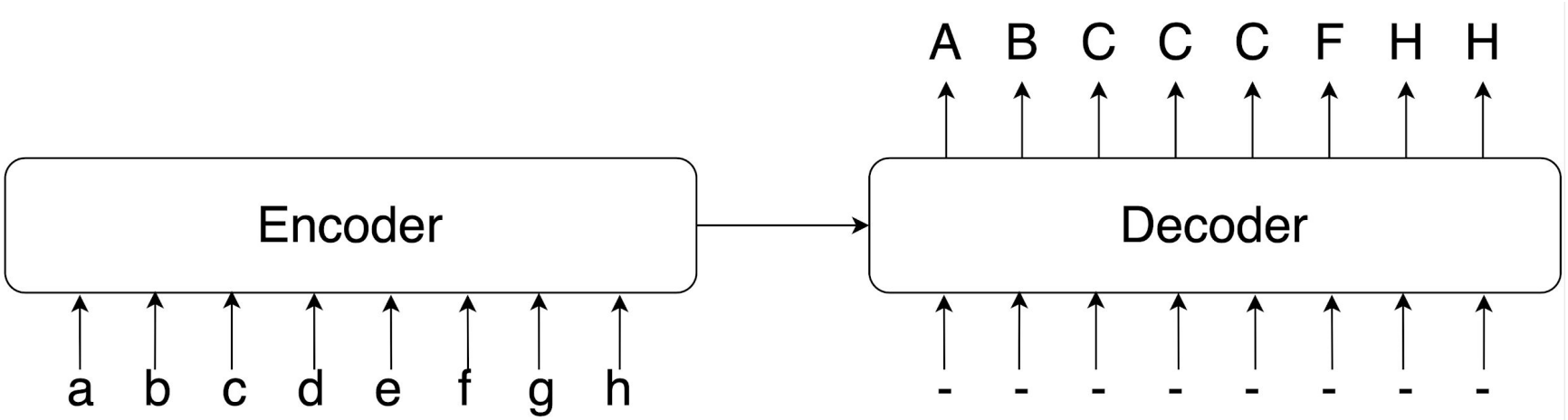
Архитектура

Авторы убирают ограничение декодера на просмотр будущих токенов





Multimodality problem



Multimodality problem

Thank you

Danke

Danke schön

Vielen Dank

Неверные переводы

Danke Dank

Vielen schön

Токены предсказания не зависят друг от друга

Решение проблемы

Запустим алгоритм несколько раз на каждом шаге будем предсказывать токены в которых алгоритм был наименее уверен на предыдущем шаге

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

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

Замаскированные токены

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

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

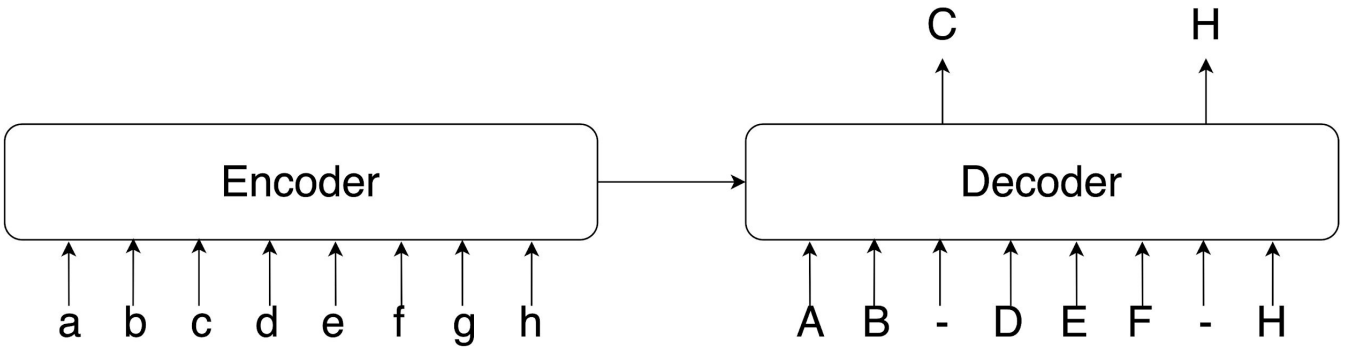
Остальные токены

$$y_i^{(t)} = y_i^{(t-1)}$$

$$p_i^{(t)} = p_i^{(t-1)}$$

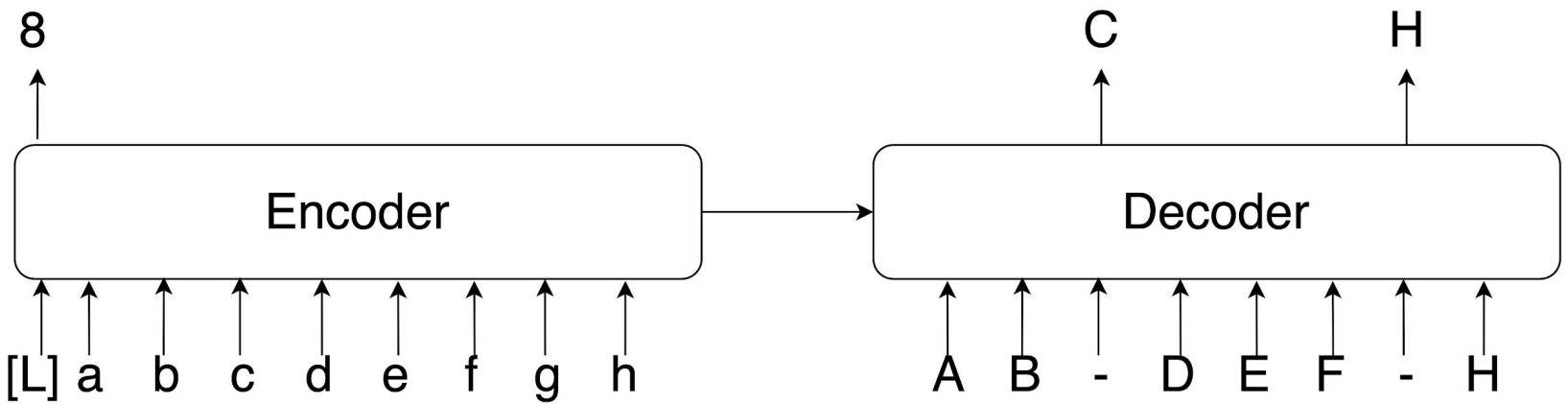
T - количество итераций
t - номер итерации

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



Обучение модели

- Трансформер без ограничения на просмотр токенов справа
- Маскируем k ~ Unif(1, N) токенов
- Предсказываем замаскированные токены
- Предсказываем длину последовательности



Предсказание длины последовательности

- добавим токен длины к энкодеру
- токен длины энкодера моделирует распределение длин последовательностей
- можем взять l наиболее вероятных длин энкодера
- предсказать для них перевод
- взять лучший перевод

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

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	—

Итоговый алгоритм

Инициализация

- предсказываем длину
- присваиваем всем таргетам значение <mask>
- предсказываем все токены

Итерация

- маскируем наименее уверенные токены
- предсказываем их

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

Количество итераций

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%

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

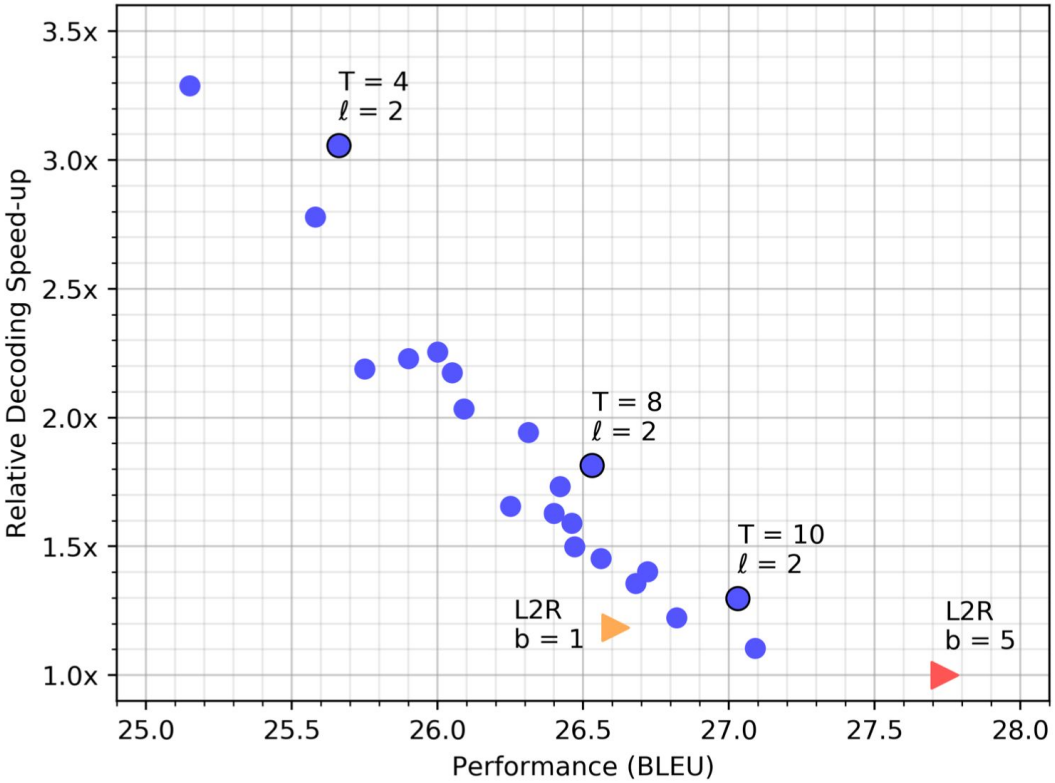
Distillation

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**

Результаты

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	— —	— —

Результаты


3.5x
3.0x
2.5x
2.0x
1.5x
1.0x
Relative Decoding Speed-up
T = 4
ℓ = 2
T = 8
ℓ = 2
T = 10
ℓ = 2
L2R
b = 1
L2R
b = 5
25.0
25.5
26.0
26.5
27.0
27.5
28.0
Performance (BLEU)

Источники

1. Non-Autoregressive Neural Machine Translation https://arxiv.org/pdf/1711.02281.pdf
2. Attention Is All You Need https://arxiv.org/pdf/1706.03762.pdf
3. Mask-Predict: Parallel Decoding of Conditional Masked Language Models https://arxiv.org/pdf/1904.09324.pdf

Вопросы

1. В чем заключается Multimodality problem, приведите пример?
2. Какие токены маскируются на i-ой итерации алгоритма?
3. Приведите схему Parallel Decoding of Conditional Masked Language Models.