Неавторегрессивный перевод

Сипаров Иван

Подходы к переводу текста

Авторегрессивный перевод:

$$p_{\mathcal{AR}}(Y|X;\theta) = \prod_{t=1}^{T+1} p(y_t|y_{0:t-1}, x_{1:T'};\theta)$$

$$\mathcal{L}_{\text{ML}} = \log p_{\mathcal{AR}}(Y|X;\theta) = \sum_{t=1}^{T+1} \log p(y_t|y_{0:t-1}, x_{1:T'};\theta)$$

Неавторегрессивный перевод:

$$p_{\mathcal{NA}}(Y|X;\theta) = p_L(T|x_{1:T'};\theta) \cdot \prod_{t=1}^{T} p(y_t|x_{1:T'};\theta)$$

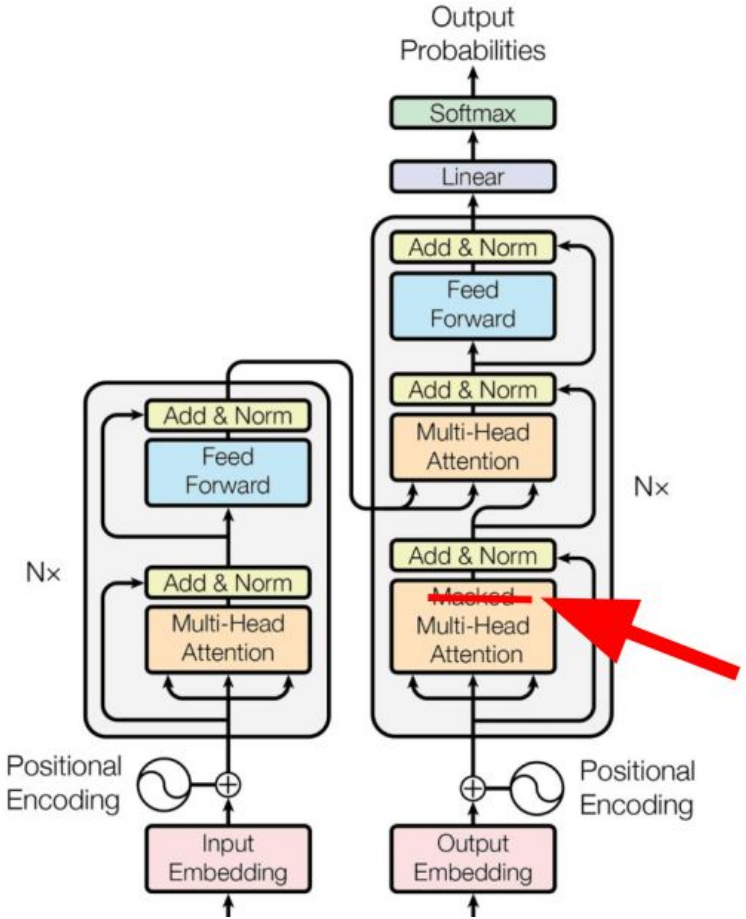
Проблемы

- Отсутствует взаимосвязь между словами
- Не понятно как выбирать длину
- Низкое качество перевода

Mask-Predict

Архитектура

- За основу взяли обычной трансформер.
- Убрали mask в декодере.



Обучение модели

- k ~ Uniform[1, N]
- Маскируем k случайно выбранных токенов
- Предсказываем замаскированные токены
- В качестве целевой функции используем кросс-энтропию
- Используем специальный токен <Length> для предсказание длины

Работа модели

1. Предсказываем N.
2. Маскируем все токены.
3. Предсказываем все токены.
4. Маскируем менее вероятные токены.
5. Предсказываем их.
6. Повторяем шаги 4-5.

Пример

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 .

Как выбирать количество итераций и маскированные токены?

- Количество итераций:
  - Константное: 1, …, 10
  - log N, √N , N

- Число маскированных токенов:

$$n = \left(1 - \frac{current\ iteration}{total\ iterations}\right) N$$

Нужно ли длинным последовательностям больше итераций?

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

Другой подход выбора длины

- Берем несколько более вероятных длин и генерируем для них последовательности
- Выбираем лучшую по формуле:

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

Влияние количества кандидатов длины на качество

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	—

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	— —	— —

Качество перевода

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

Необходима ли дистилляция?

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**

Скорость перевода