XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang et al.
Speaker: Fedor Noskov

MIPT, 792

October, 2020

AR modeling vs. autoencoding

В случае с авторегрессионной языковой моделью логарифм правдоподобия:

$$\log p_\theta(\boldsymbol{x}) = \sum_{t=1}^{T} \log p_\theta(x_t | \boldsymbol{x}_{<t}), \tag{1}$$

т.е. модель для предсказания нового слова смотрит только на предыдущие:

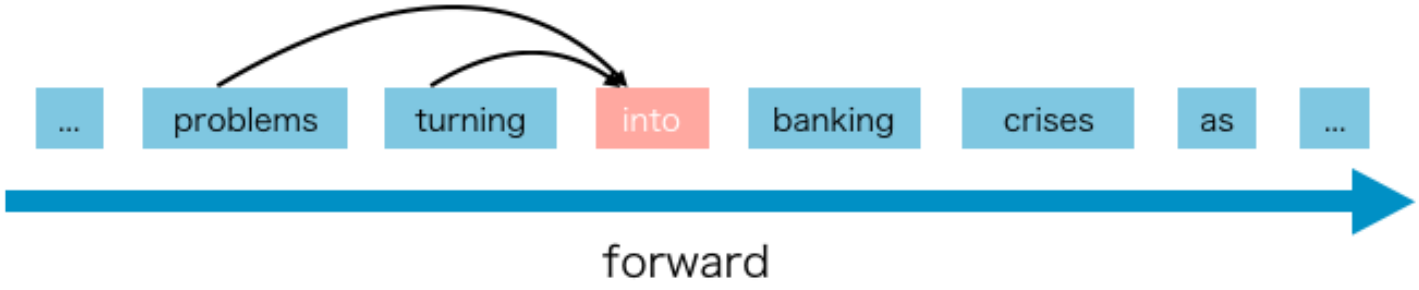

Рис. 1: (source)

AR modeling vs. autoencoding

В случае BERT предполагаем, что при условии некоторого контескта $\hat{\boldsymbol{x}}$ остальные токены появляются независимо:

$$\log p_\theta(\boldsymbol{x}_{masked}|\hat{\boldsymbol{x}}) = \sum_{t=1}^{T} \mathrm{I}\{x_t \text{ is unknown}\} \log p_\theta(x_t|\hat{\boldsymbol{x}}) \qquad (2)$$

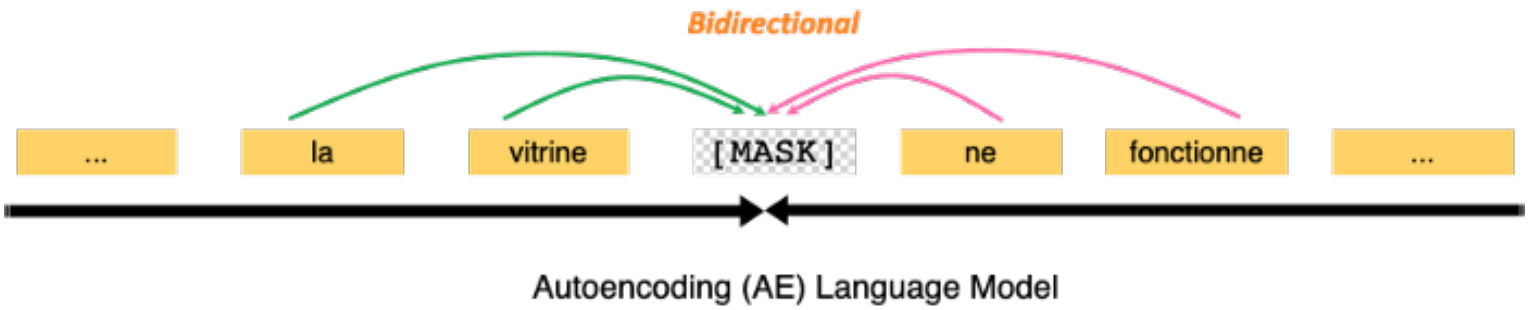

Рис. 2: (source)

AR modeling vs. autoencoding

Но в реальности это предположение неверно! Решение: заметим, что порядок обуславливания неважен.

1)	___	???	___
2)	___	немедленно	???
3)	???	немедленно	выпил
4)	И	немедленно	выпил

Т.е. для любой перестановки $\boldsymbol{z}$:

$$\log p_\theta(\boldsymbol{x}) = \sum_{t=1}^{T} \log p_\theta(x_{z_t} | \boldsymbol{x}_{<\boldsymbol{z}\, t}) \tag{3}$$

AR modeling vs. autoencoding

Будем максимизировать следующую величину:

$$\mathbb{E}_{\boldsymbol{z}\sim\mathcal{Z}_T} \sum_{t=1}^{T} \log p_\theta(x_t|\boldsymbol{x}_{<_{\boldsymbol{z}}t}) \tag{4}$$

А параллельность?..

Two-Stream Self-Attention

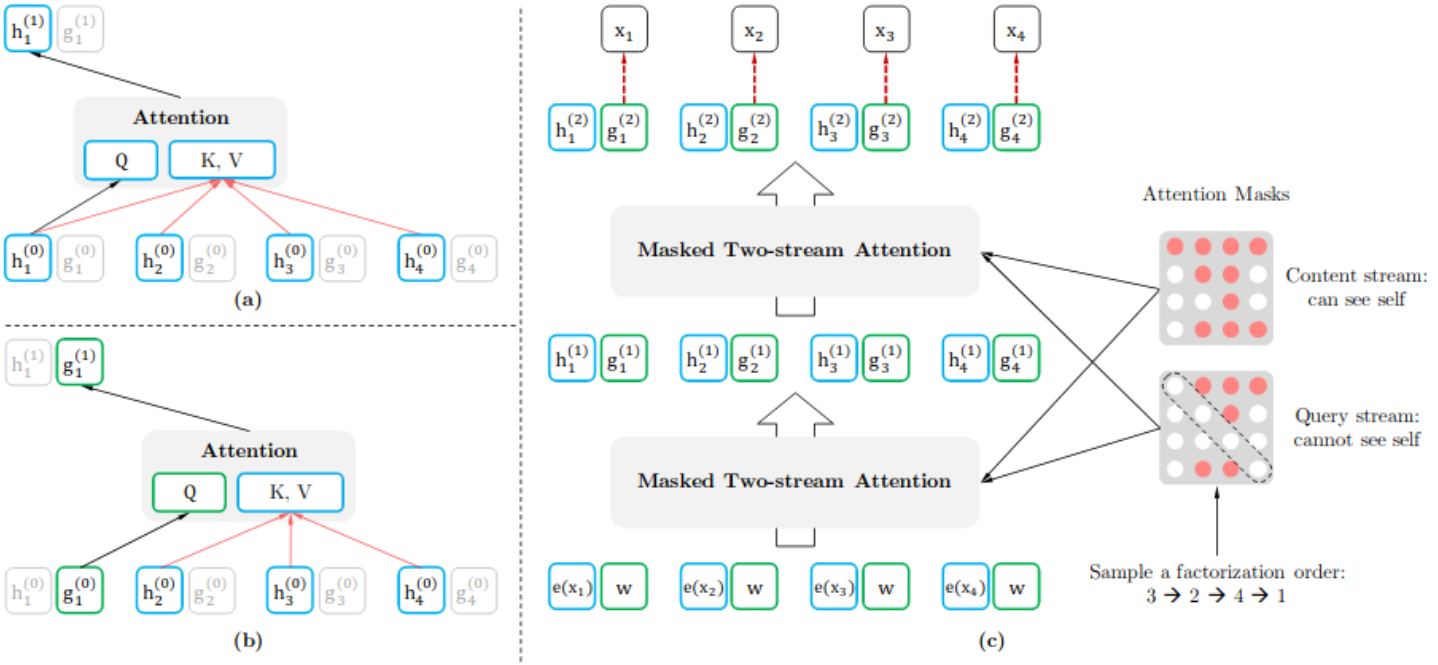

Рис. 3: (source)

Тонкости обучения

- Обучать с учетом всех перестановок сложно. По этой причине будем в последовательности предсказывать только два последних токена
- Переставлять длинную последовательность сложно. Будем разбивать ее на несколько частей и для каждой части передавать в слой трансформера вектор контекста предыдущих. Перестановки будем совершать только внутри этих частей
- Стоимость обучения около 245 тыс. долларов

Результаты

Model	IMDB	Yelp-2	Yelp-5	DBpedia	AG	Amazon-2	Amazon-5
CNN [15]	-	2.90	32.39	0.84	6.57	3.79	36.24
DPCNN [15]	-	2.64	30.58	0.88	6.87	3.32	34.81
Mixed VAT [31, 23]	4.32	-	-	0.70	4.95	-	-
ULMFiT [14]	4.6	2.16	29.98	0.80	5.01	-	-
BERT [35]	4.51	1.89	29.32	0.64	-	2.63	34.17
XLNet	**3.20**	**1.37**	**27.05**	**0.60**	**4.45**	**2.11**	**31.67**

Рис. 4: Классификационный датасеты