Language models are unsupervised task learners

Долженко Николай
НИУ ВШЭ

Задача

- Большинство современных датасетов в NLP направлены на какую-то конкретную задачу и/или область.

- Часто используется обучение с учителем

- Плохо ведут себя на документах из другого распределения

- Языковые модели учатся предсказывать условную вероятность:

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1})$$

- Это позволяет сэмплировать распределение вида:

$$p(s_{n-k}, ..., s_n | s_1, ..., s_{n-k-1})$$

- Т.е. P(*output*| *input*)

- Хотим считать распределения вида:

P(*output*| *input, task)*

В силу того, что мы работаем с языком обучающая выборка может иметь вид не (task, input, output), а быть представлена одной последовательностью символов

Например в задаче перевода объект обучающей выборки может быть записан как:

“translate to russian; I like cats; Мне нравятся кошки”

В такой постановке, задачи NLP, требующие обучающую выборку (перевод, ответы на вопросы), являются частным случаем моделирования языка.

Основная идея статьи состоит в том, что если языковую модель обучать на большом корпусе разнообразных текстов, то она также будет обучаться различным задачам, которые встречаются в языке, чтобы лучше предсказывать последовательности.

Сбор данных

- Большинство обучающих выборок для языковых моделей содержат тексты из одной области (газеты, художественная литература)
- Поставленная задача требует много осмысленных текстов из различных областей.
- Существующие датасеты собранные в интернете часто имеют много документов с бессвязным текстом

WebText

- Люди хорошо оценивают качество документов
- Reddit — платформа с огромным количеством ссылок на документы
- Взяты все ссылки встреченные на reddit из постов с хорошей оценкой
- Итоговый датасет, после того как убрали дубликаты, имеет размер 40GB и содержит более 8 миллионов документов.

Представление текста

- Модель должна потенциально уметь предсказывать любую unicode строку.

- Многие современные модели используют предобработку текста (перевод в нижний регистр, токенизацию, замена слов, которых нет в словаре специальным токеном)

- Второй вариант представлять текст последовательностью байт. Такие модели в теории могут работать, но на практике показывают себя хуже, чем высокоуровневое представление текста

Byte Pair Encoding

- Токенезировать текст не по словам, а по часто встречающимся последовательностям байт.
- При этом также иметь токены для всех символов
- Из-за жадного алгоритма построения представления, многие слова могут быть плохо объединены. Поэтому запрещают объединять символы из разных классов (пунктуацию и буквы).
- Пример проблемы: dog. dog? dog! могут стать разными токенами

Архитектура GPT и GPT2

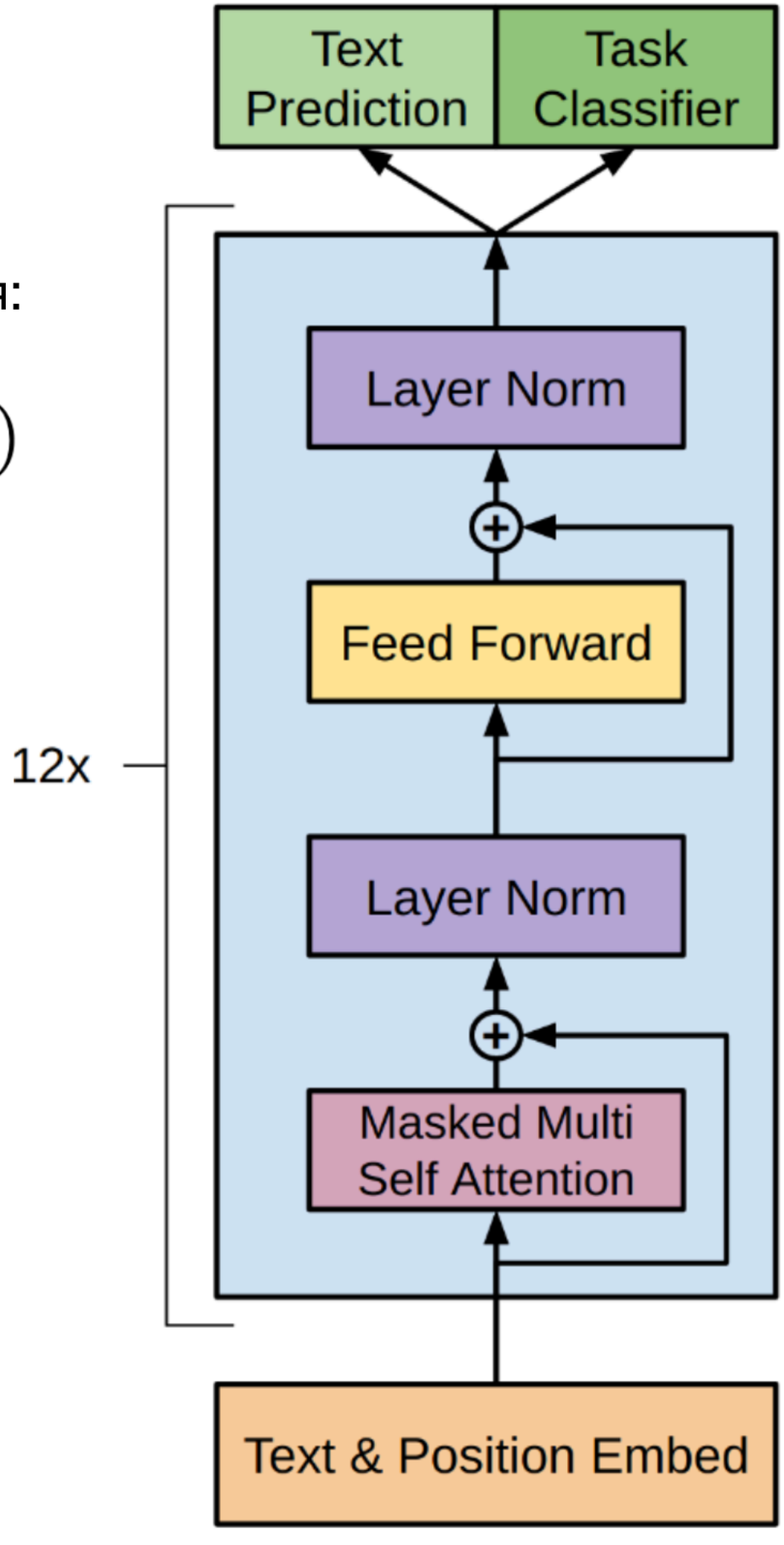
Для набора токенов {u_1, …, u_n}, максимизируется:

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

где *k* — размер контекста

- В GPT2 размер контекста 1024 (512 в GPT)
- Количество токенов 50257
- И размера батча 512

Всего в GPT2 1542 миллиона параметров



Эксперименты

Language modeling

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

GPT2 показывает себя значительно лучше State of the art моделей на многих тестах.

Summarization

	R-1	R-2	R-L	R-AVG
Bottom-Up Sum	**41.22**	**18.68**	**38.34**	**32.75**
Lede-3	40.38	17.66	36.62	31.55
Seq2Seq + Attn	31.33	11.81	28.83	23.99
GPT-2 `TL;DR:`	29.34	8.27	26.58	21.40
Random-3	28.78	8.63	25.52	20.98
GPT-2 no hint	21.58	4.03	19.47	15.03

Для того, чтобы получить краткую сводку текста из GPT2 генерируется 100 токенов и из них берутся первые три законченных предложения.
На каждой итерации берется случайный из двух наиболее вероятных токенов

Для того, чтобы “стимулировать” GPT2 генерировать краткую сводку, к тексту в конце добавляется строка TL;DR: (too long, didn’t read)

Перевод

Качество перевода оценивалось на тестах WMT-14 english-french и WMT-14 french-english

Для того, чтобы модель начала генерировать перевод предложения, перед ним добавлялось несколько строк вида: французское предложение = английское и после переводимого предложения ставилось “=“

С английского на французский модель набирает 5 BLEU, что немного хуже, чем просто замена переводом каждого слова

С французского на английский: 11.5 BLEU, что лучше некоторых современных моделей, которые обучаются без учителя, но хуже рекорда в 33.5 BLEU.

Проблемы с существующими датасетами

	PTB	WikiText-2	enwik8	text8	Wikitext-103	1BW
Dataset train	**2.67%**	0.66%	**7.50%**	2.34%	**9.09%**	**13.19%**
WebText train	0.88%	**1.63%**	6.31%	**3.94%**	2.42%	3.75%

Пример генерируемого текста

Context (human-written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

GPT-2: The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Вопросы

- **Авторы статьи проводили анализ пересечения документов в различных датасетах. Какой метод описывается в статье для определения пересечения в документах?**
- **Какое представление данных (токенизацию) используют авторы статьи?**
- **В статье авторы описывают применение General Language Model в задачах обощения (summarization) и машинного перевода. Какое изменение текста приводится в статье, чтобы при inference на нем модель вернула его перевод?**

Источники

- Language Models are Unsupervised Multitask Learners. Radford et. al

- Improving Language Understanding by Generative Pre-Training. Radford et. al