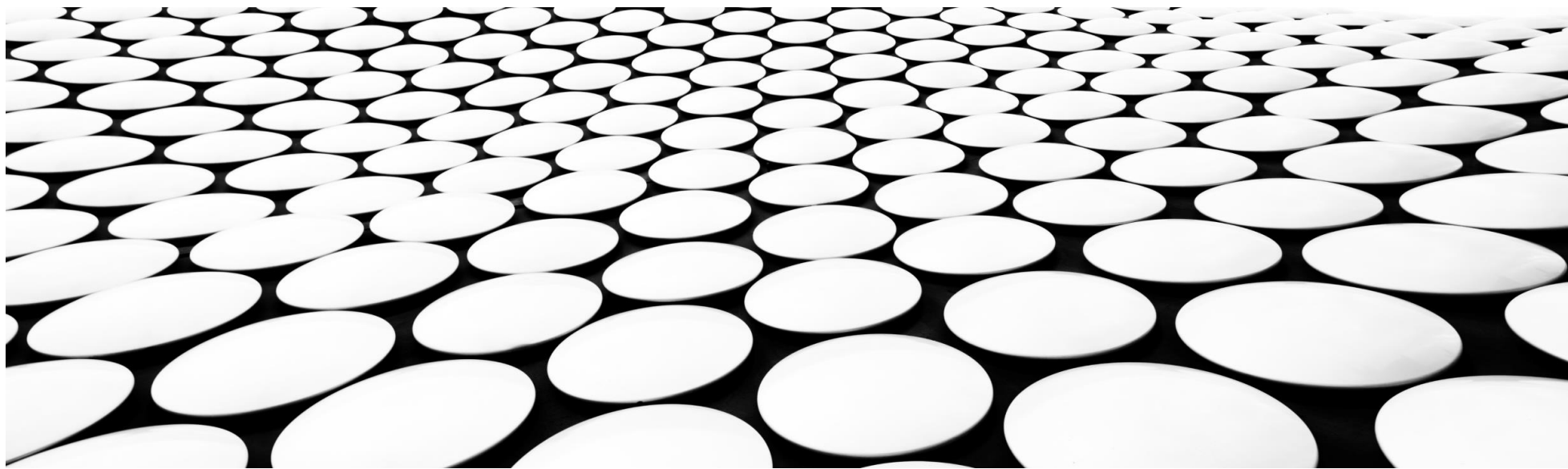


深度学习

邱怡轩



今天的主题

- 注意力机制与 Transformer
- 大语言模型初探



The New York Times

Wednesday, June 9, 2010 1:42 PM EDT



Search

Change Home Layout

Facebook

Twitter

Google+

Subscribe to Home Delivery | Translate Our Website

Switch to
Mobile Edition

ARTS
BUSINESS
COLUMNISTS

WORLD
U.S.
POLITICS
NEW YORK
BUSINESS
DEADLINE
TECHNOLOGY
SPORTS

SCIENCE
HEALTH
OPINION
ARTS

Books
Movies
Music
Television
Theater
Style
Dining & Wine
Fashion & Style
Home & Garden
Real Estate
Travel

ALL ABOUT
Calendar
Classifieds
Community
Cooking & Food
Education
Find Love
Lifestyle
Multimedia
NYC Guide
Obituaries
Personal
Public Safety
Real Estate
Marketing

Cuomo Urges Broad Limits to N.Y. Public Pensions

By David L. Musto and Thomas H. Johnson 2:45 PM EDT
Gov. Andrew M. Cuomo said that knowing other changes, New York State and City should pass the retirement age for new public employees to 55.

Full & Complete Read (17)

Pick for Afghan Envoy Says U.S. Can't Afford to Abandon Effort

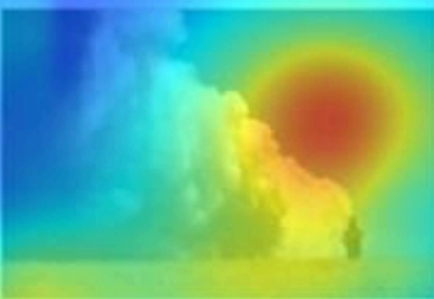
By Thomas H. Johnson 1:45 PM EDT
Ryan C. Crocker, President Obama's choice for envoy to Afghanistan, said the United States could not afford to walk away despite the cost and spotty progress.

Full & Complete Read (16)

Banks Defeated in Senate Vote Over Debit Card Fees

By Thomas H. Johnson 4:41 PM EDT
The Senate rejected a delay in regulations over debit card fees, essentially leaving it to the Federal Reserve to limit the fees that stores pay banks.

Full & Complete Read (16)



Firefighters Struggle With Arizona Wildfire

By Thomas H. Johnson 4:32 PM EDT
A massive fire continued its surge and west burning earlier today, igniting dozens of smaller fires.



Iran Plans High Level of Uranium Enrichment

By Thomas H. Johnson 1:35 PM EDT
Iran declared that it aims to triple production of nuclear fuel this year and increase enrichment to 20 percent.

Full & Complete Read (16)

OPEC Keeps Lid on Oil Production Targets

By Thomas H. Johnson 2:45 PM EDT
Over the objections of Saudi Arabia, OPEC on Wednesday left quoted as place despite rising world prices.

Full & Complete Read (17)

OPINION
- **Don't Blame the Spread!**
How we might avoid another brightening of the U.S. economy.

- Freshman: The Earth Is Full
- Dwell: 'Sour Tastes' of the Heart
- Editorial: Healthcare Crisis
- Op-Ed: The Gas Is Greener
- Dispatch: The Switzerland of America

Log In With Facebook

Log in and see what your friends are sharing on Twitter.com. Privacy Policy / What's New?

WHAT'S POPULAR NOW

- The Earth Is Full
- Obama Congressman Is Observed in the Cam, but Aide Says He'll Run for Re-election - New York Times

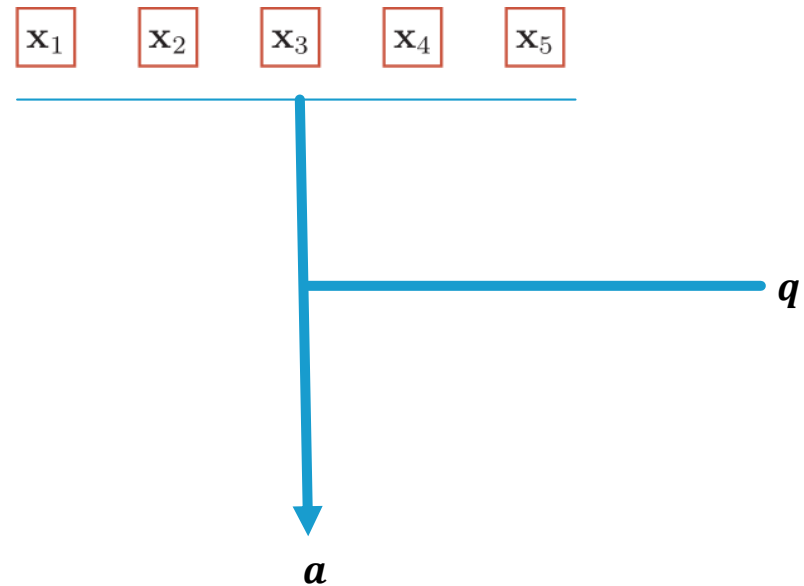
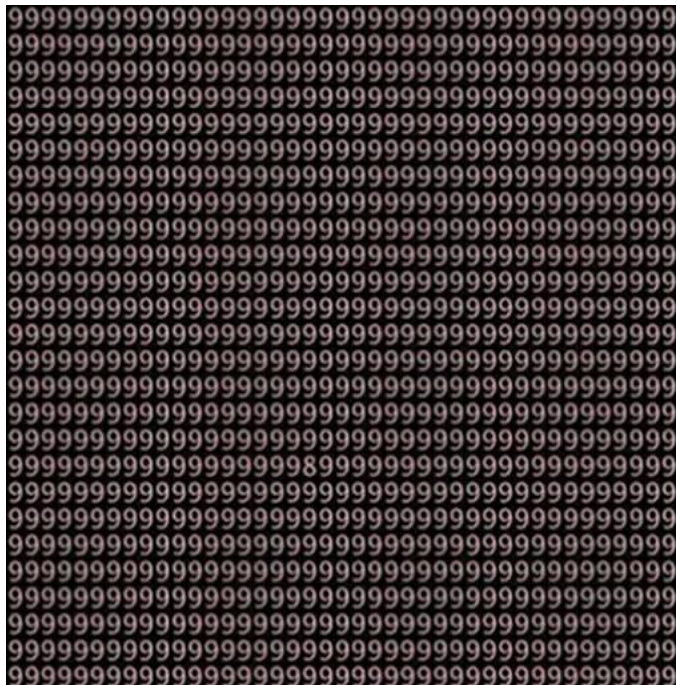


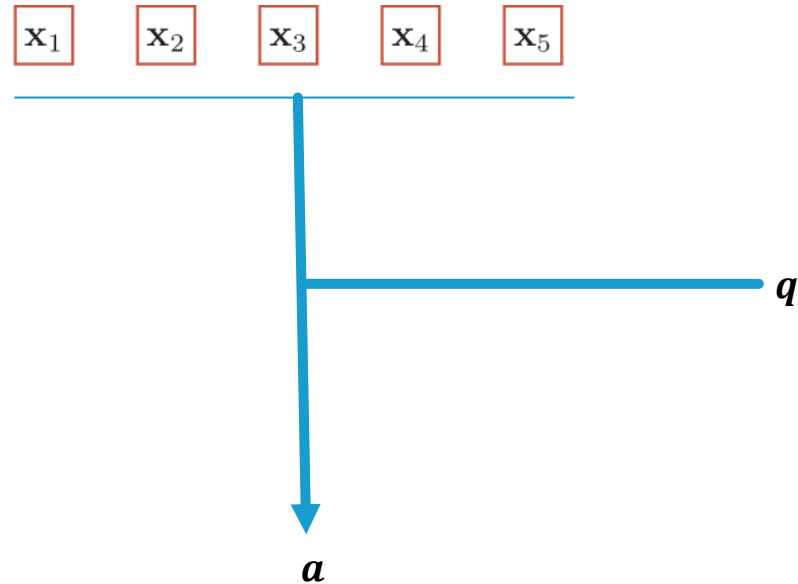
MARKETS

	5:28 PM	Down	Up
S&P 500	1,279.34	-12.44	2,471.17
DAX	4,211.87	-42.17	2,471.17
NASDAQ	2,471.17	-24.71	2,471.17
10Y Treasury	107.0101	-0.0101	107.0101
Gold	1,279.34	-12.44	2,471.17

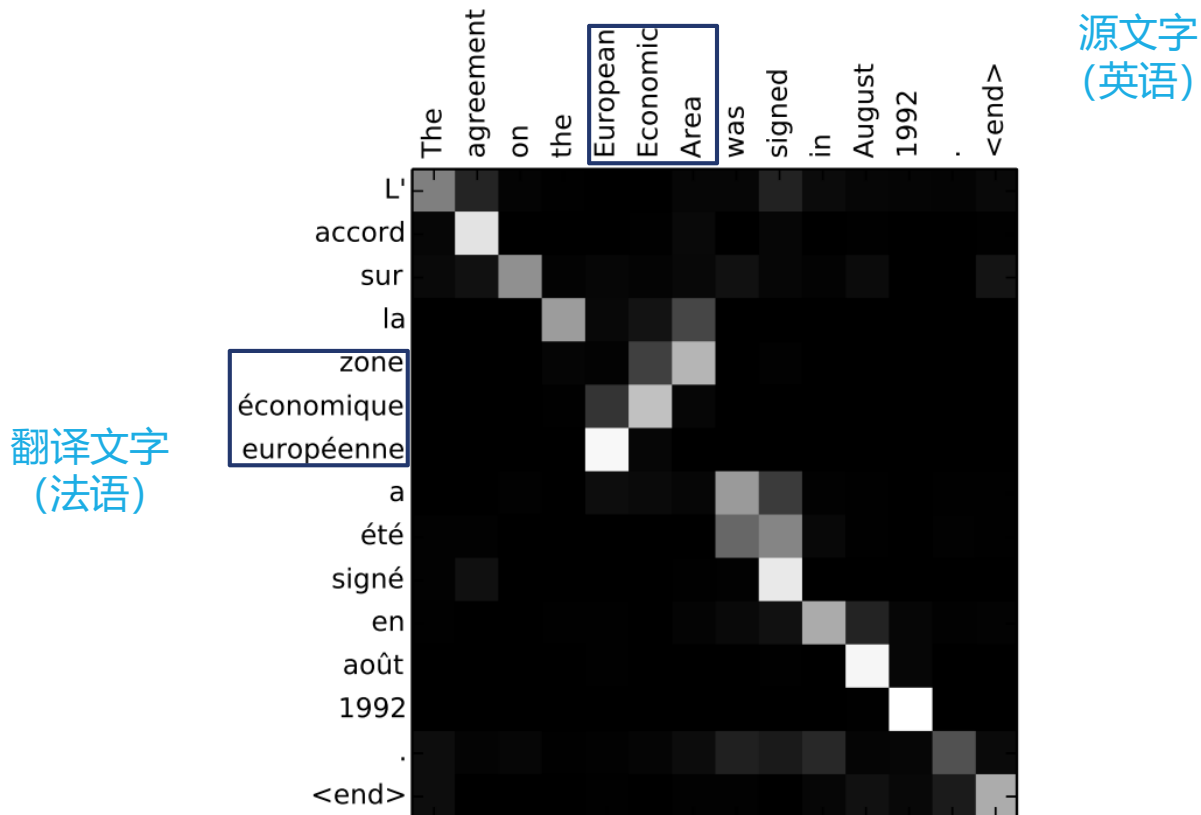
TIMES DIGITAL SUBSCRIPTIONS
We have a special offer for you. Get the New York Times digital edition for \$10.99 a month. [Click here to subscribe.](#)





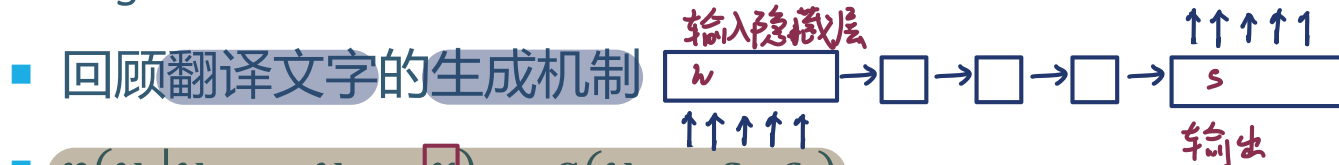


- 在机器翻译中，注意力机制可以理解作为一种文字“对齐”的方法



实现方法

- 基于 *Neural Machine Translation by Jointly Learning to Align and Translate*. ICLR 2015.



- $p(y_i | y_1, \dots, y_{i-1}, \boxed{x}) = g(y_{i-1}, s_i, c_i)$

↑↑↑↑↑

输入

- y_i 是当前翻译出的单词

- s_i 是 RNN 当前的隐藏层向量, $c_i = s_{i-1}$ 与原来相同, $s_i = f(s_{i-1}, y_{i-1}, c_i)$

- c_i 是当前上下文向量, 相当于对输入序列进行压缩后的结果

- 注意力机制体现在 c_i 的选择和构建上

实现方法

- 整体而言, c_i 是对输入序列隐藏层的加权平均
- $c_i = \sum_{j=1}^T \alpha_{ij} h_j$
- 权重 α_{ij} 代表了第 j 个输入文字对当前翻译输出的“重要性”
- $(\alpha_{i1}, \dots, \alpha_{iT}) = \text{softmax}(e_{i1}, \dots, e_{iT})$
- $e_{ij} = a(s_{i-1}, h_j)$ 是一个打分函数, 参数数量固定

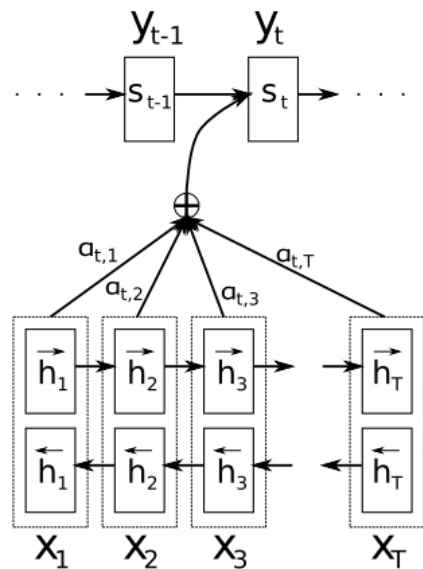


Figure 1: The graphical illustration of the proposed model trying to generate the t -th target word y_t given a source sentence (x_1, x_2, \dots, x_T) .

打分函数

- 打分函数有不同的形式
- 但核心在于参数数量固定，不随输入序列长度影响

加性模型

$$s(\mathbf{x}, \mathbf{q}) = \mathbf{v}^\top \tanh(\mathbf{W}\mathbf{x} + \mathbf{U}\mathbf{q}),$$

点积模型

$$s(\mathbf{x}, \mathbf{q}) = \mathbf{x}^\top \mathbf{q},$$

缩放点积模型

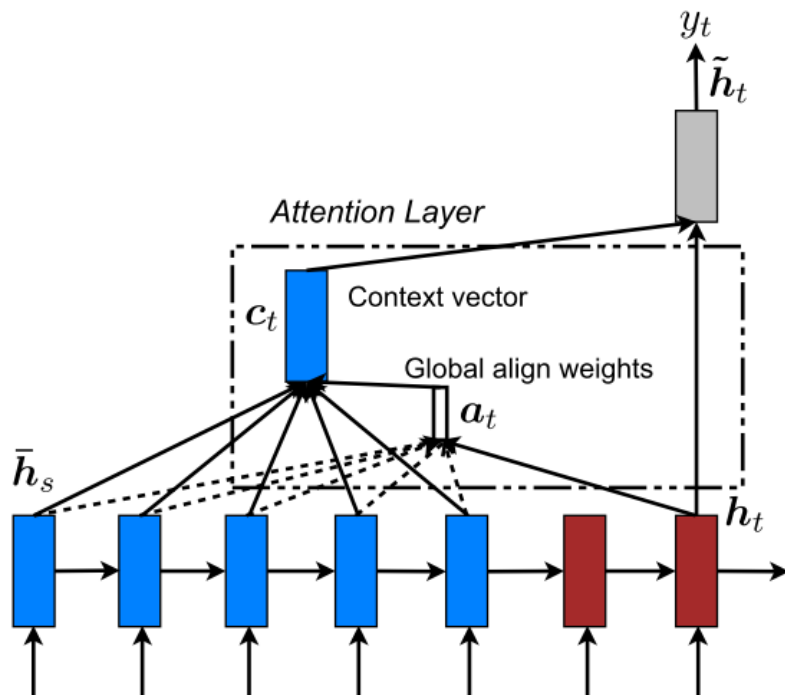
$$s(\mathbf{x}, \mathbf{q}) = \frac{\mathbf{x}^\top \mathbf{q}}{\sqrt{D}},$$

双线性模型

$$s(\mathbf{x}, \mathbf{q}) = \mathbf{x}^\top \mathbf{W} \mathbf{q},$$

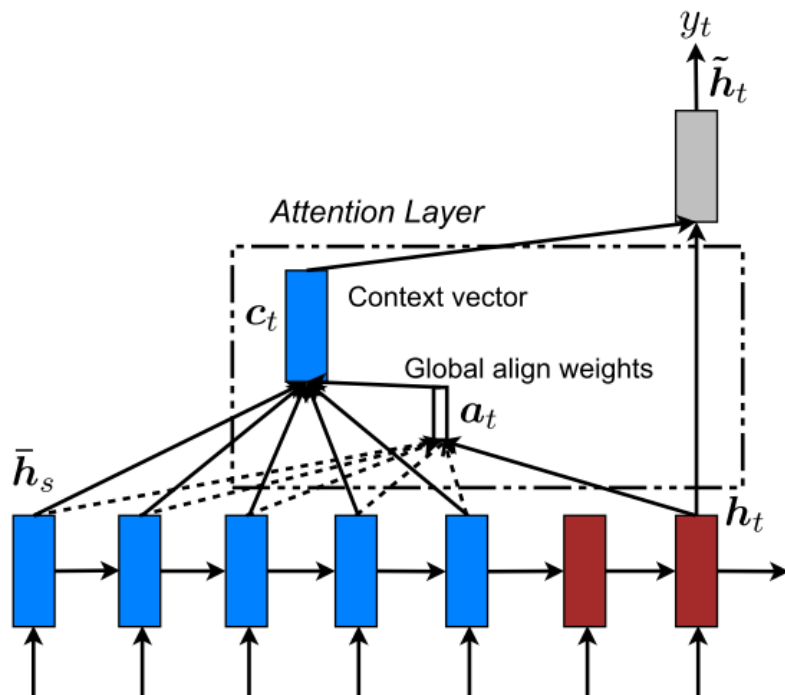
全局注意力

- Effective Approaches to Attention-based Neural Machine Translation 一文对上述结构做了细微改动，并提出“全局注意力”的概念



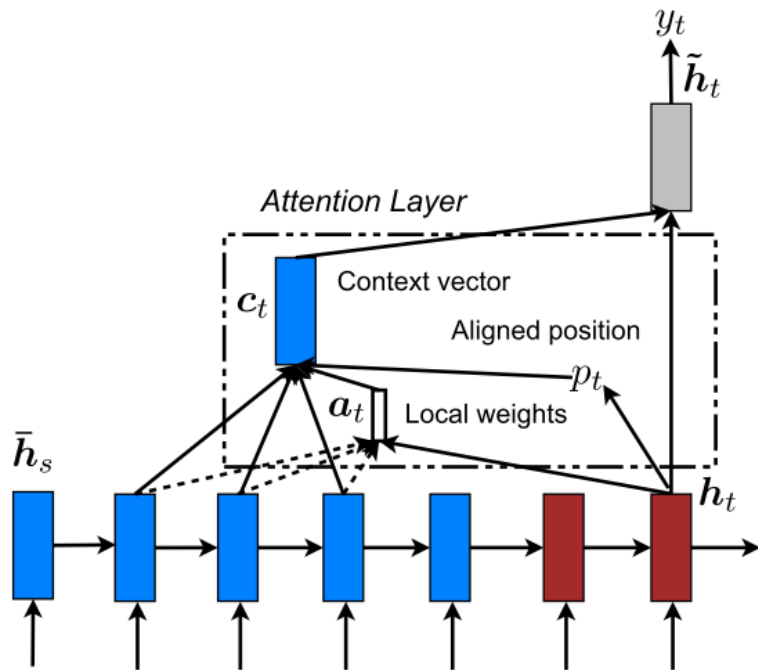
全局注意力

- “全局”的含义在于，生成上下文向量 c_t 时利用到了所有输入元素的隐藏层向量



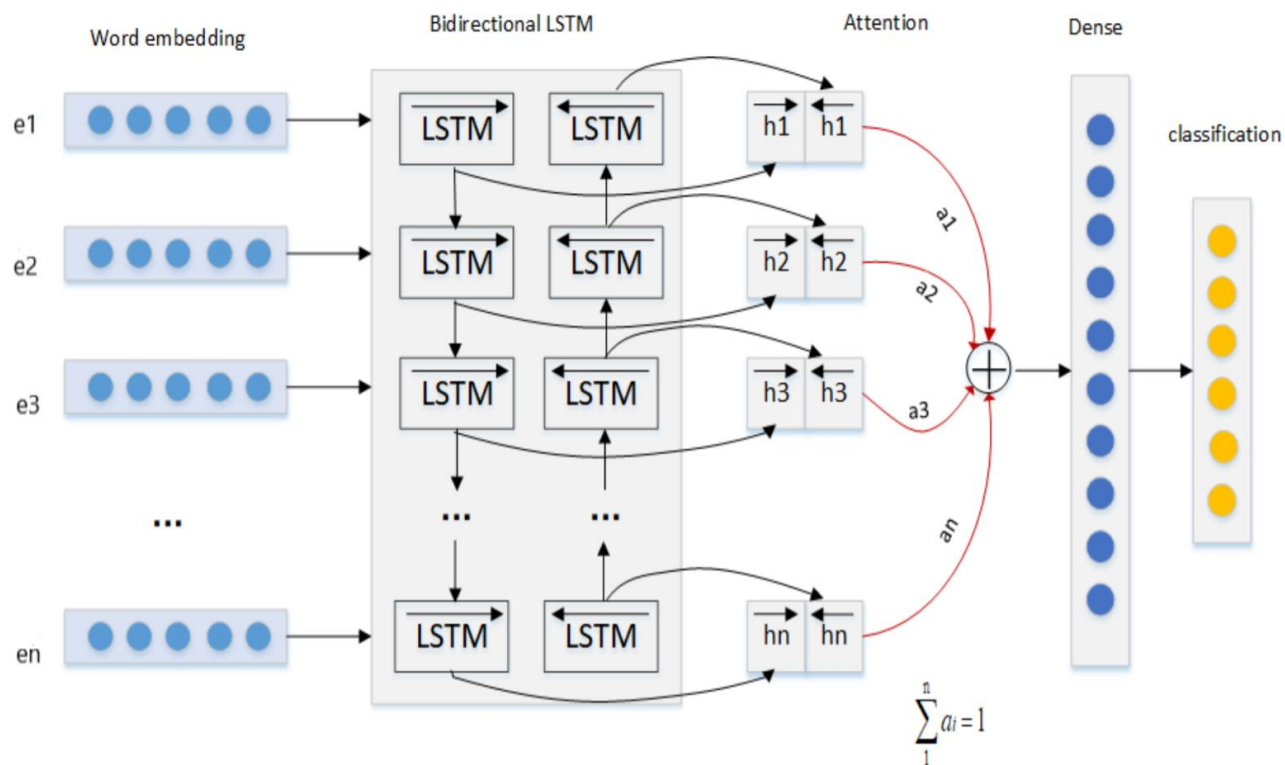
局部注意力

- 与全局相对的，是局部注意力
- a_t 计算的范围是一个固定宽度的滑动窗口
- 好处在于减少了计算量



其他应用

■ 文本分类



其他应用

- 图像标注
- 给定文字，标识出图片对应的区域



A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



A giraffe standing in a forest with trees in the background.



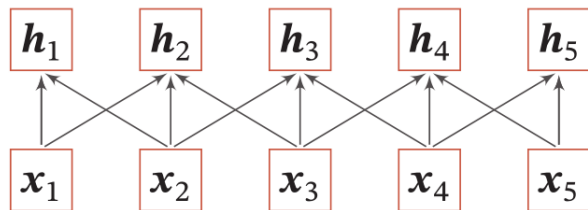
Attention Is All You Need (?)

Attention 热潮

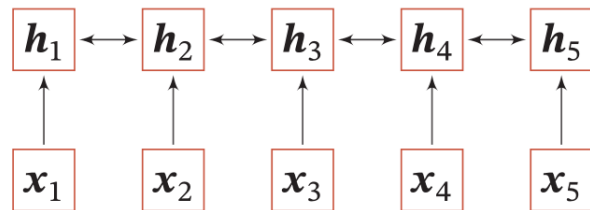
- 自注意力 (Self-attention)
- 多头注意力 (Multi-attention)
- KQV模式 (Key-Query-Value)
- Transformer
- BERT
- GPT
- Vision Transformer
-

自注意力

- 自注意力模型进一步放松了 RNN 的限制
- RNN 的主要作用在于利用固定数量的参数将不定长的序列映射为对应的隐藏表示
- 卷积网络也有类似的效果



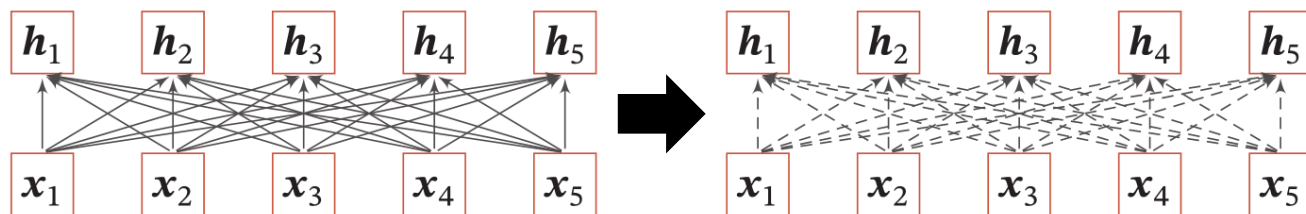
(a) 卷积网络



(b) 双向循环网络

自注意力

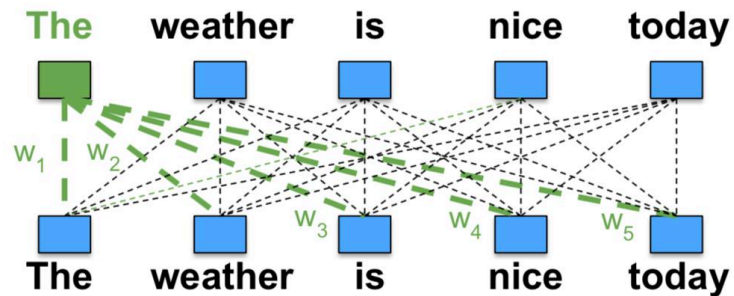
- 然而 RNN 和卷积主要利用的是局部信息
- 全连接网络可以跨越较长的距离，但参数数量不固定
- 自注意力模型可以看作是一种特殊的全连接层，权重由注意力机制生成，参数数量固定



(a) 全连接模型

(b) 自注意力模型

自注意力



$$w_1, w_2, w_3, w_4, w_5 = \text{softmax} \left(\begin{bmatrix} 0.6 & 0.2 & 0.8 \end{bmatrix} \times \begin{bmatrix} 0.6 & 0.2 & 0.9 & 0.4 & 0.4 \\ 0.2 & 0.3 & 0.1 & 0.1 & 0.1 \\ 0.8 & 0.1 & 0.8 & 0.4 & 0.6 \end{bmatrix} \right)$$

The The weather is nice today

$$\begin{bmatrix} 1.8 \\ 2.3 \\ 0.4 \end{bmatrix} = w_1 \times \begin{bmatrix} 0.6 \\ 0.2 \\ 0.8 \end{bmatrix} + w_2 \times \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \end{bmatrix} + w_3 \times \begin{bmatrix} 0.9 \\ 0.1 \\ 0.8 \end{bmatrix} + w_4 \times \begin{bmatrix} 0.4 \\ 0.1 \\ 0.4 \end{bmatrix} + w_5 \times \begin{bmatrix} 0.4 \\ 0.1 \\ 0.6 \end{bmatrix}$$

The The weather is nice today

KQV 模式

- 注意力机制还可以进一步抽象
- 输入序列的隐藏表示由三部分组成
- Key-Query-Value

KQV 模式

- 回顾
- 上下文向量 $c_i = \sum_{j=1}^T \alpha_{ij} h_j$
- 权重 $(\alpha_{i1}, \dots, \alpha_{iT}) = \text{softmax}(e_{i1}, \dots, e_{iT})$
- 打分函数 $e_{ij} = a(s_{i-1}, h_j)$

KQV 模式

- 回顾

隐藏表示是 Value 的加权平均

- 上下文向量 $c_i = \sum_{j=1}^T \alpha_{ij} h_j$

- 权重 $(\alpha_{i1}, \dots, \alpha_{iT}) = \text{softmax}(e_{i1}, \dots, e_{iT})$

- 打分函数 $e_{ij} = a(s_{i-1}, h_j)$

将 Query 去和 Key 进行匹配

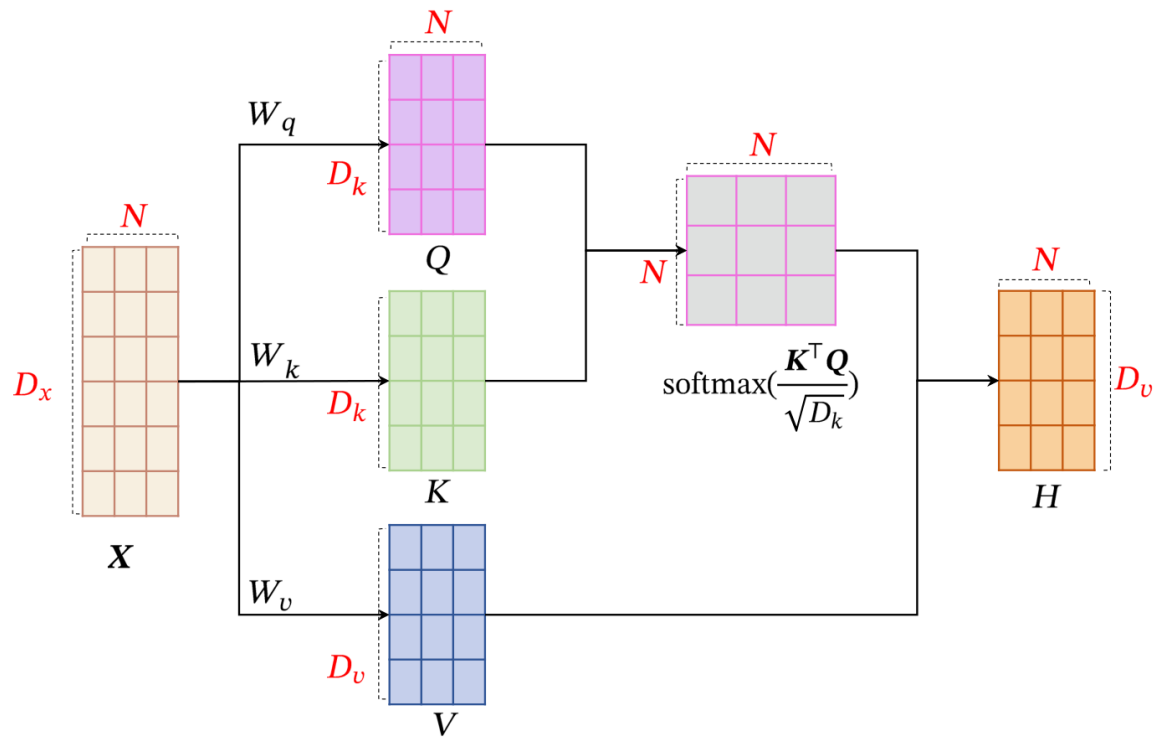
KQV 模式

$$Q = W_q X \in \mathbb{R}^{D_k \times N},$$

$$K = W_k X \in \mathbb{R}^{D_k \times N},$$

$$V = W_v X \in \mathbb{R}^{D_v \times N},$$

$$H = V \operatorname{softmax}\left(\frac{K^\top Q}{\sqrt{D_k}}\right),$$



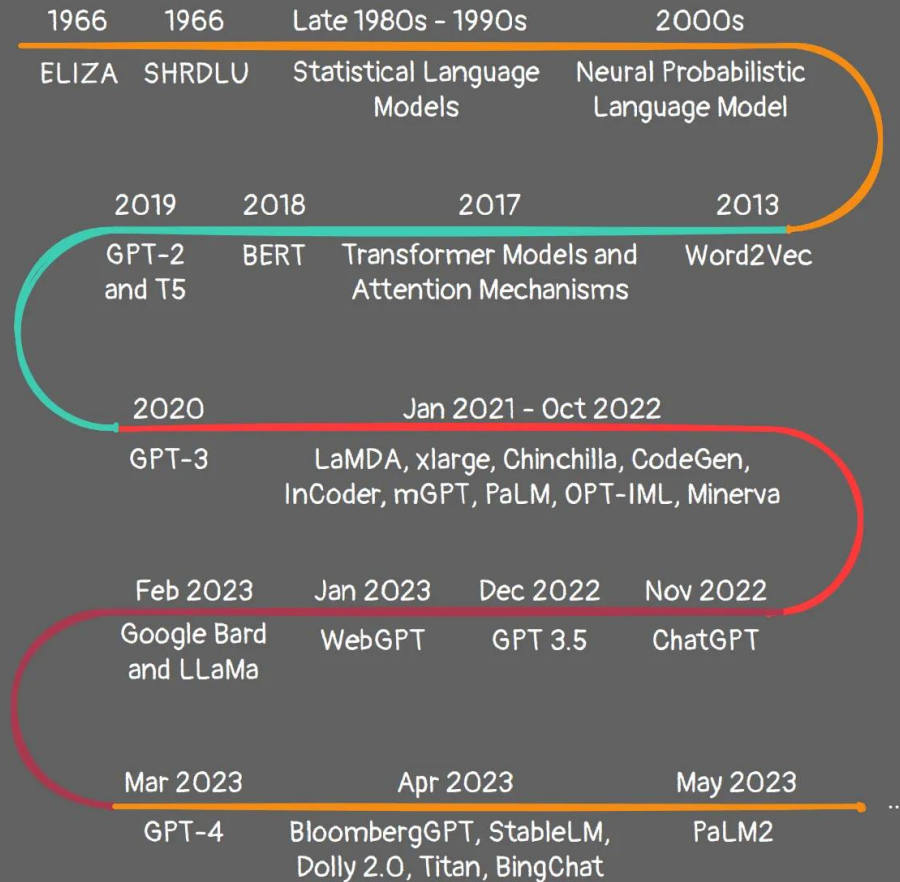
思考

- 为什么要想尽办法计算隐藏表示？
- 为了更好地利用数据和问题的结构
- 不同的模型架构可能理论表达能力是等价的
- 但实际中要根据数据的特征进行设计

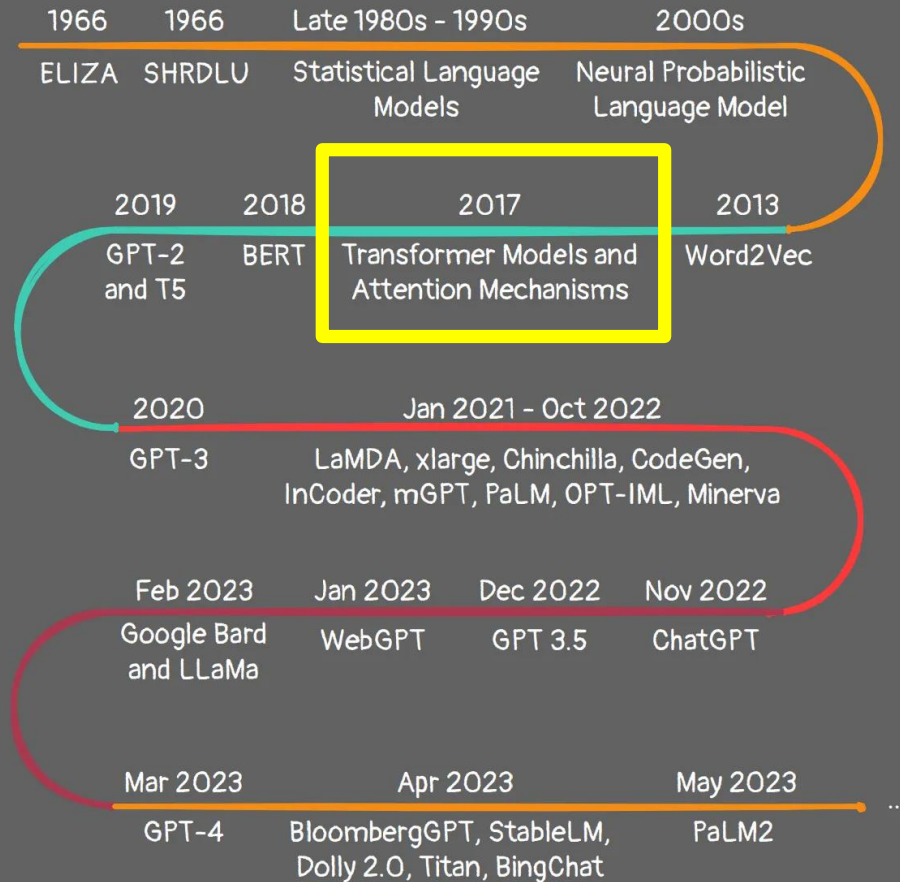


大语言模型初探

The brief history of Large Language Models



The brief history of Large Language Models



Transformer

- 2017年, 一篇名为 *Attention is all you need* 的文章标志着 Transformer 横空出世
- Transformer 可以看成是利用 Attention 来设计的一种网络结构
- 成为当今几乎所有大语言模型的基本构成单位

Transformer

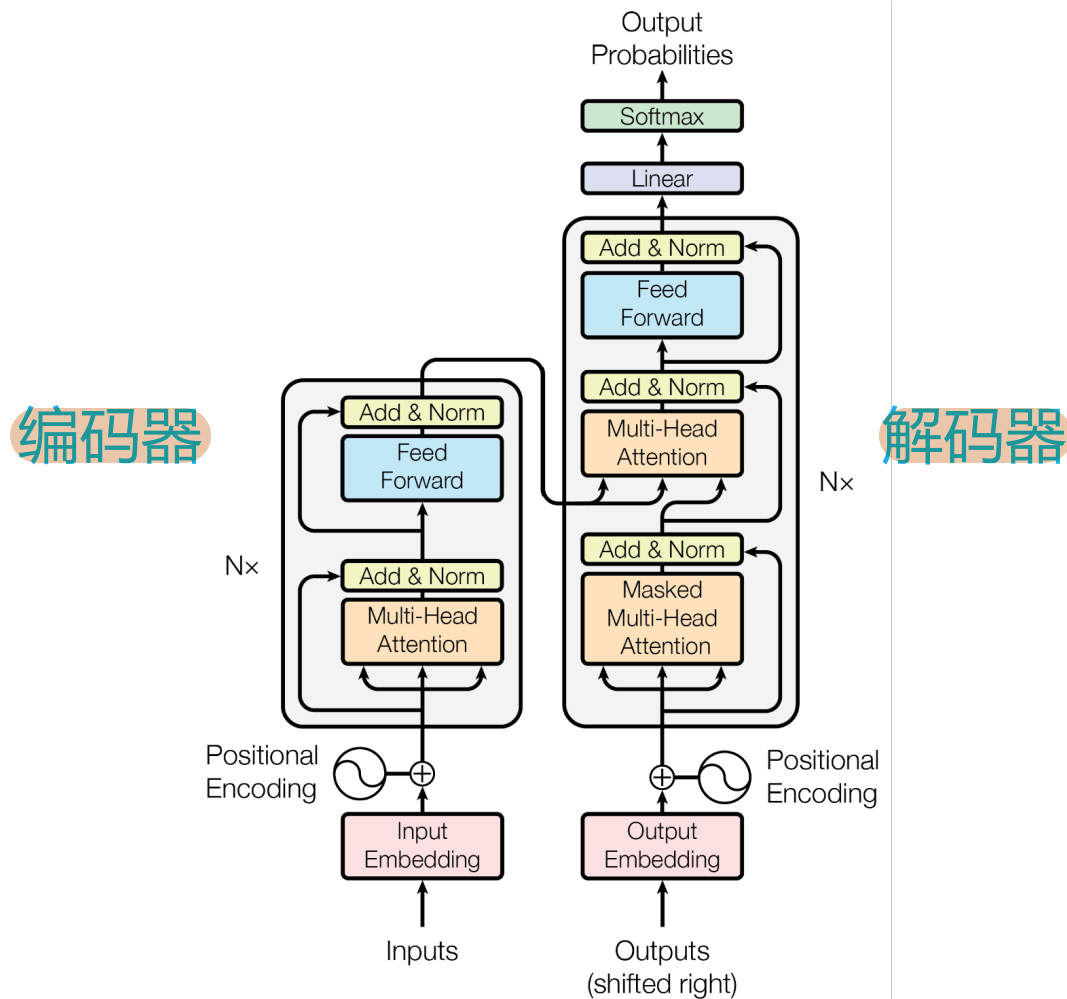
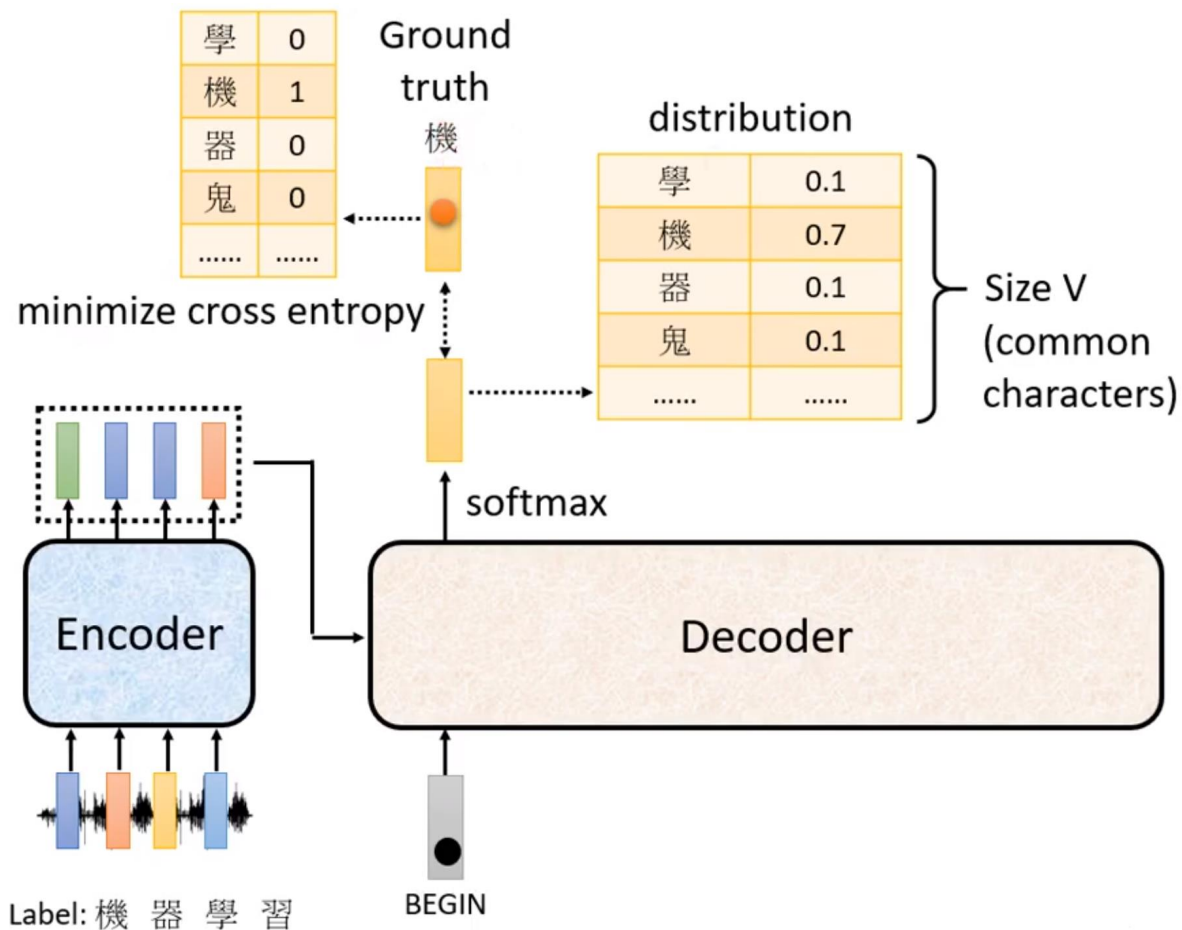


Figure 1: The Transformer - model architecture.

Transformer

本页图片取自李宏毅
Transformer 讲义



Transformer

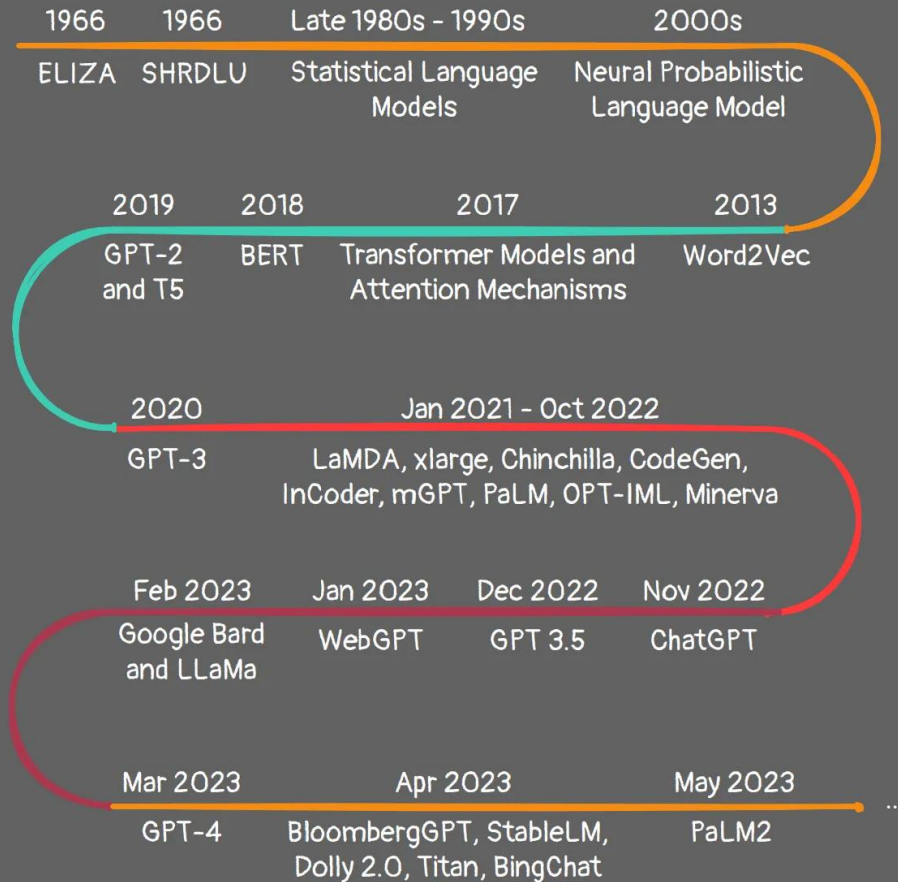
- Transformer 的影响力之大，使其甚至进入了一些流行文化和影视作品之中



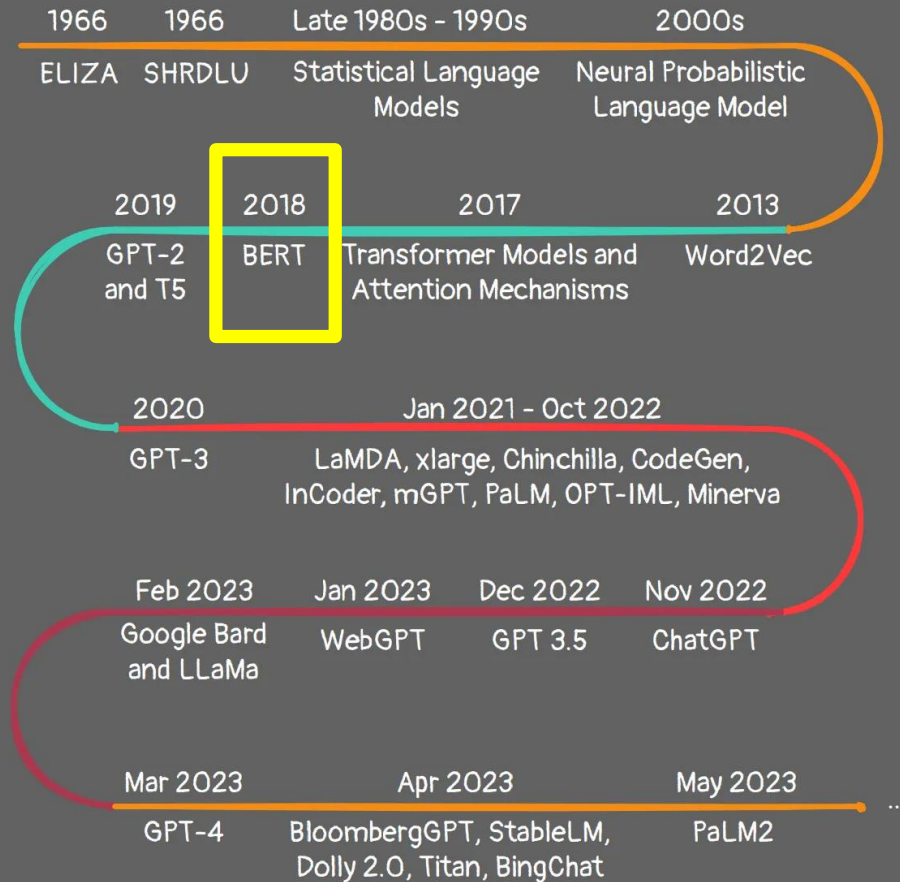
扩展阅读

- <https://www.bilibili.com/video/BV1pu411o7BE/>

The brief history of Large Language Models



The brief history of Large Language Models



BERT

- BERT 的全称为 Bidirectional Encoder Representations from Transformers
- 同样是利用了 Transformer 结构构建出的语言模型

BERT

- BERT 只利用了 Transformer 的编码器
- 采用自监督学习的方法
- “完形填空”：随机选择一些词语，将其盖住 (MASK)
- 自监督学习的目的就是用文本剩余的部分来预测被盖住的词语
- 使用双向信息：不一定从左读到右



BERT

本页图片取自李宏毅

BERT 讲义

Masking Input

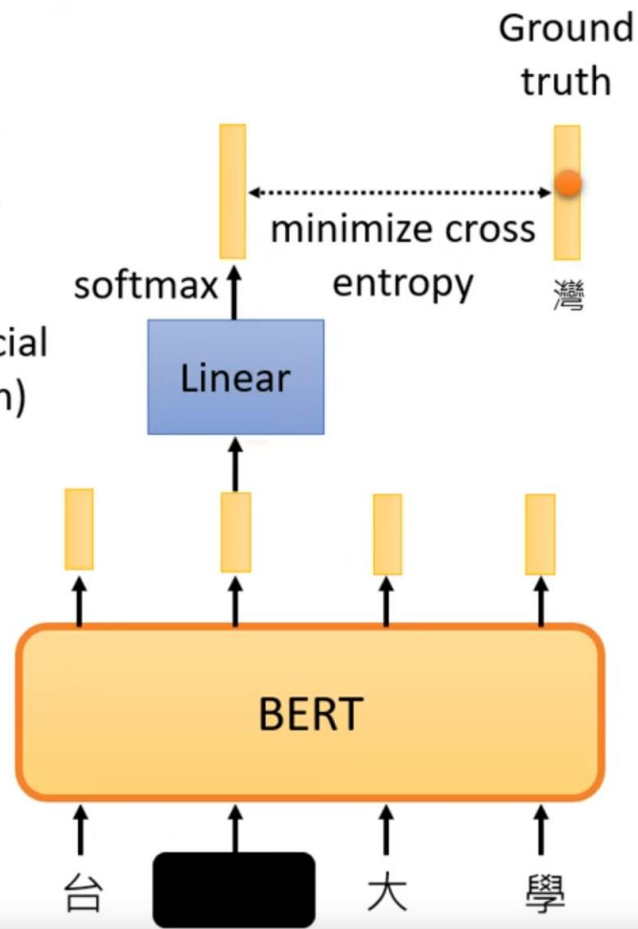
<https://arxiv.org/abs/1810.04805>

 =  (special token)
or

 = 
一、天、大、小 ...

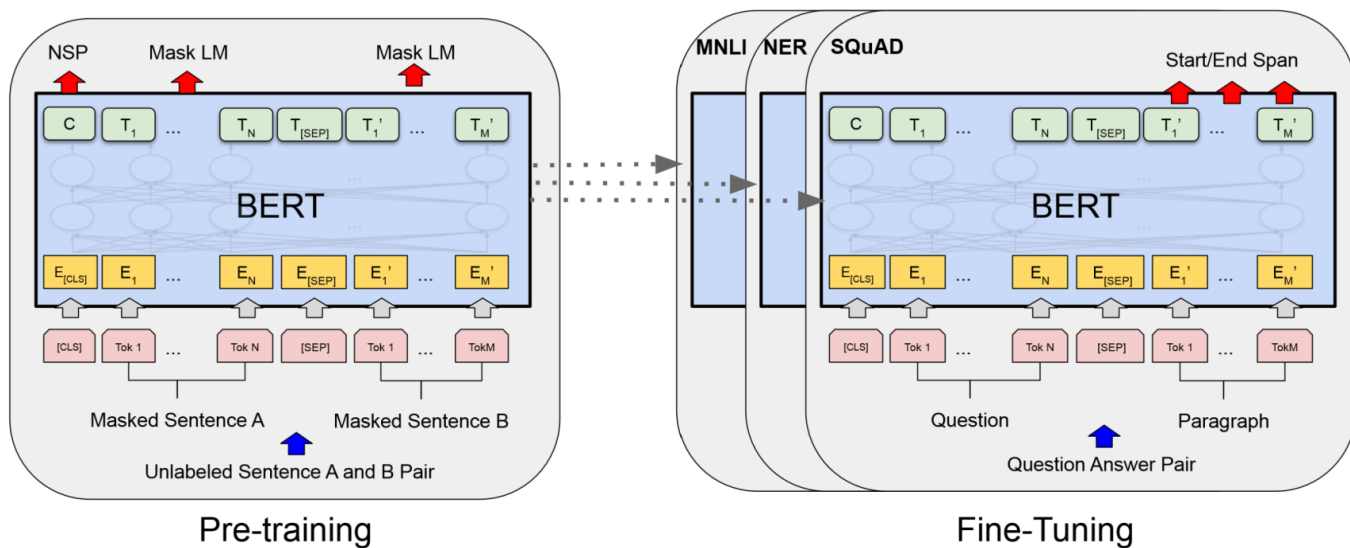
Transformer
Encoder

Randomly masking
some tokens



BERT

- BERT 极大地普及了 “预训练+微调” 这一大模型的使用模式
- 在训练阶段获取文本的嵌入向量表达
- 微调阶段用向量表达完成各项 NLP 任务

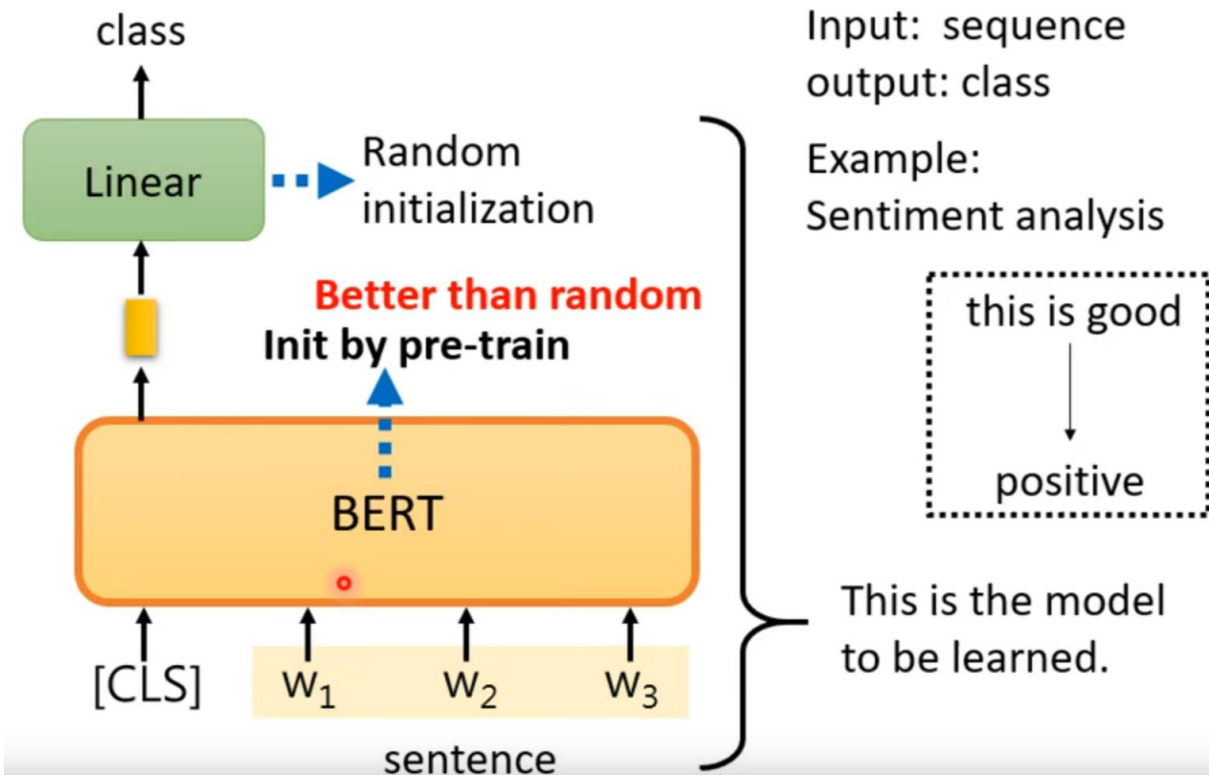


BERT

本页图片取自李宏毅

BERT 讲义

How to use BERT – Case 1

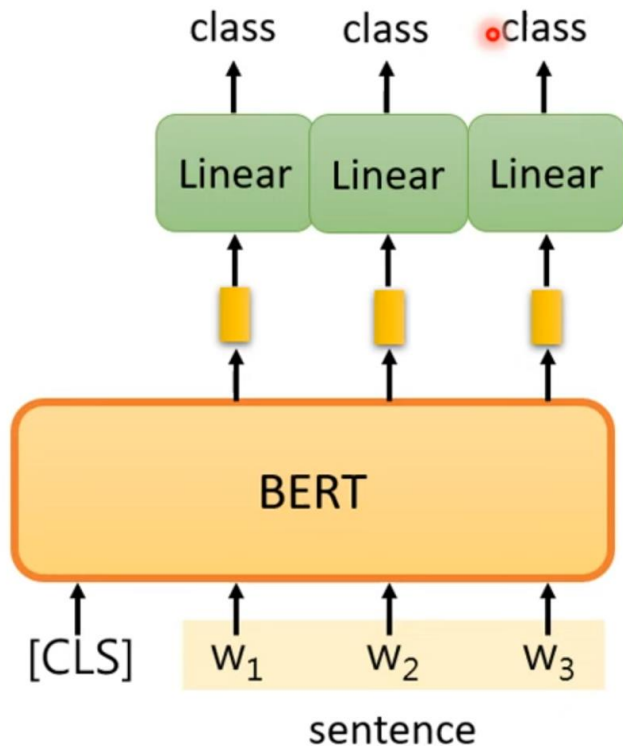


BERT

本页图片取自李宏毅

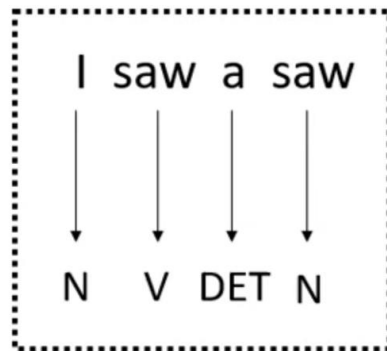
BERT 讲义

How to use BERT – Case 2



Input: sequence
output: same as input

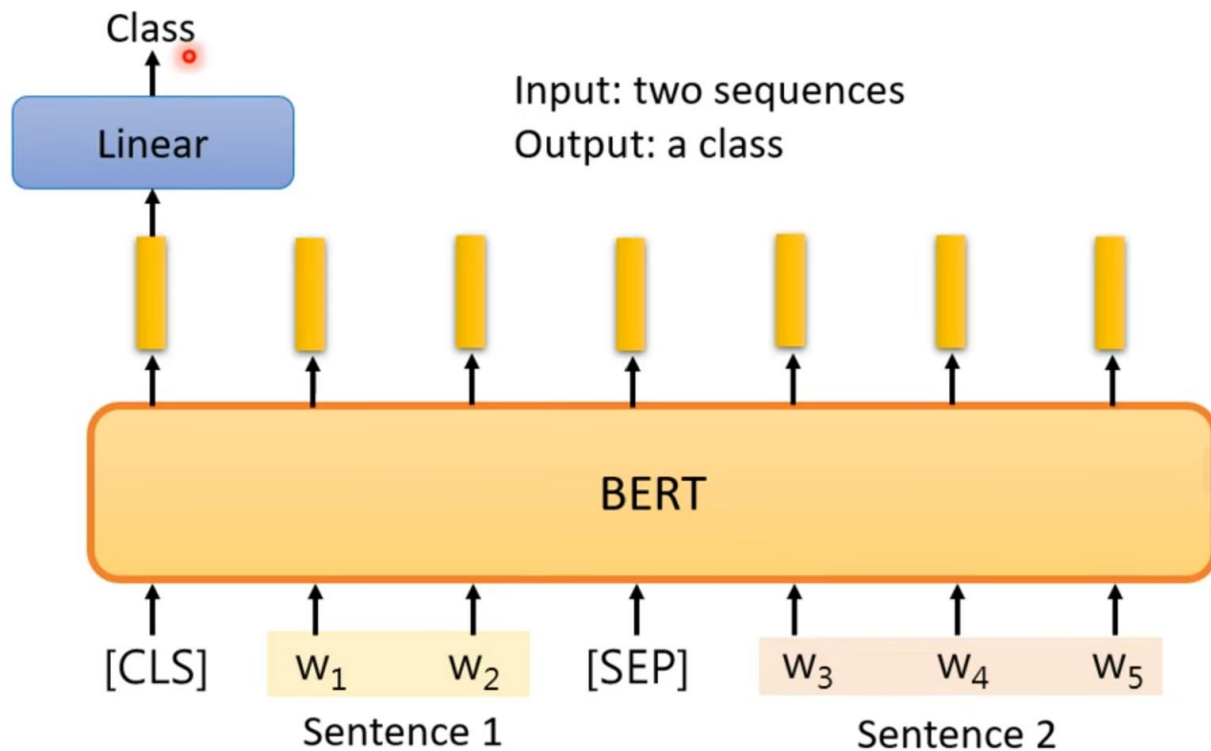
Example:
POS tagging



BERT

本页图片取自李宏毅
BERT 讲义

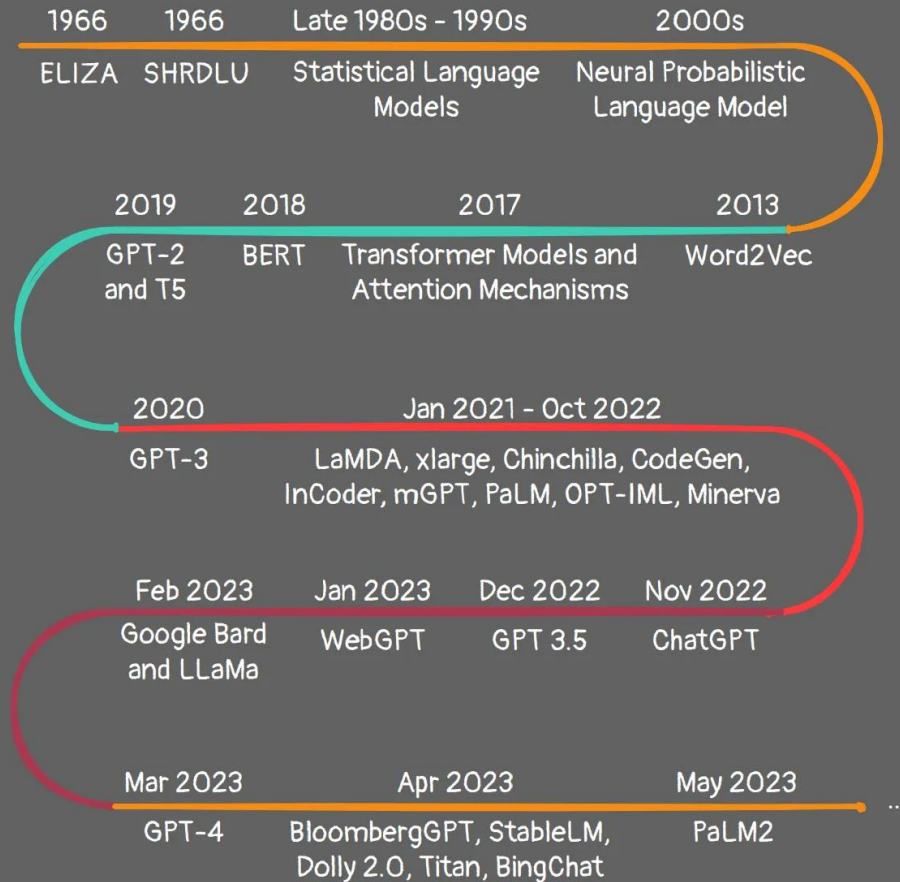
How to use BERT – Case 3



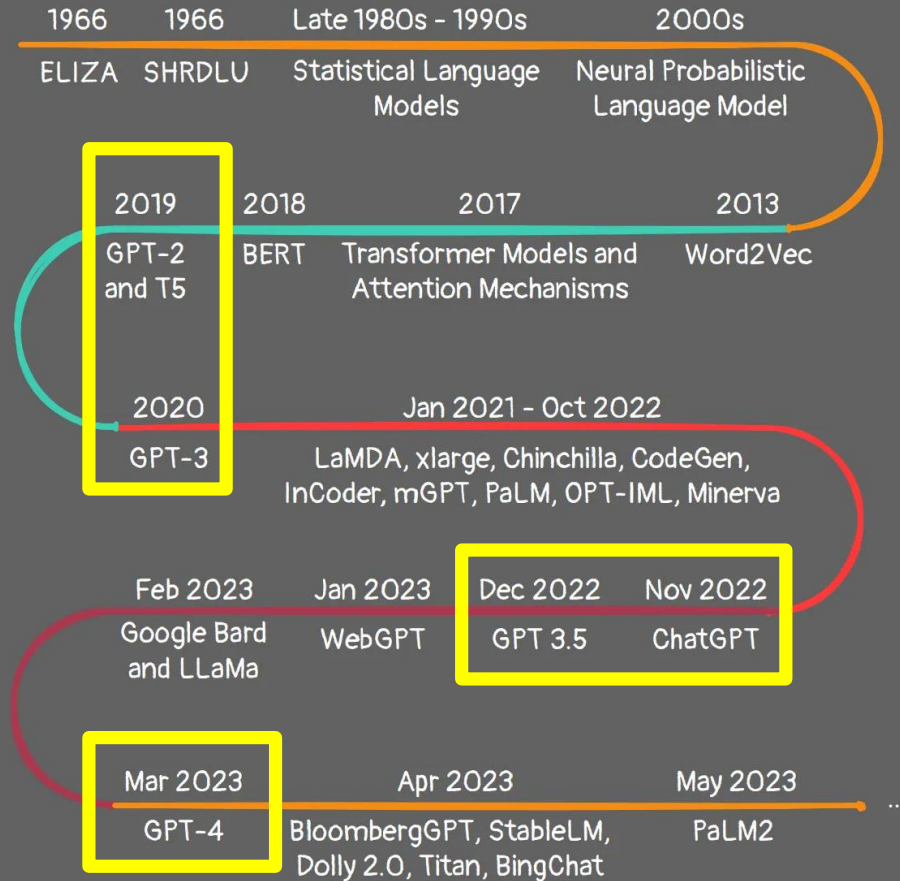
扩展阅读

- <https://www.bilibili.com/video/BV1PL411M7eQ/>

The brief history of Large Language Models



The brief history of Large Language Models



GPT

- Generative Pre-trained Transformer
- 由 OpenAI 主导开发的系列模型
- GPT-1, GPT-2, GPT-3, ChatGPT, GPT-4.....

GPT

- GPT 走上了一条与 BERT 不同的技术路线
- BERT 解决 NLP 任务的思想是“预训练+微调”
- 其中微调需要针对具体任务进行设计
- 而 GPT 的路线是用自然语言本身来对模型“下指令”
- 由此产生了以 Prompt（提示词）为代表的语言模型使用方法——把各类 NLP 任务转变成问答形式

GPT

- 从训练的角度来说
 - BERT 的核心思想是 “完形填空”
 - 而 GPT 的路线就是 “文字接龙”
-
- BERT 只利用了 Transformer 的编码器
 - GPT 只利用了 Transformer 的解码器

GPT

in out

We need to stop

We need to stop anthrop

We need to stop anthropomorph

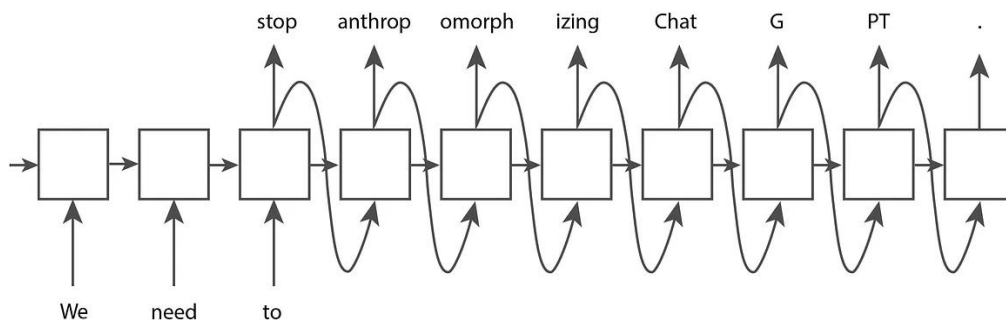
We need to stop anthropomorphizing

We need to stop anthropomorphizing Chat

We need to stop anthropomorphizing ChatG

We need to stop anthropomorphizing ChatGPT

We need to stop anthropomorphizing ChatGPT.



GPT

- 无论模型如何演变，核心的统计准则往往是朴素的

3.1 Unsupervised pre-training

Given an unsupervised corpus of tokens $\mathcal{U} = \{u_1, \dots, u_n\}$, we use a standard language modeling objective to maximize the following likelihood:

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

where k is the size of the context window, and the conditional probability P is modeled using a neural network with parameters Θ . These parameters are trained using stochastic gradient descent [51].

In our experiments, we use a multi-layer *Transformer decoder* [34] for the language model, which is a variant of the transformer [62]. This model applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

$$\begin{aligned} h_0 &= UW_e + W_p \\ h_l &= \text{transformer_block}(h_{l-1}) \forall i \in [1, n] \\ P(u) &= \text{softmax}(h_n W_e^T) \end{aligned} \quad (2)$$

where $U = (u_{-k}, \dots, u_{-1})$ is the context vector of tokens, n is the number of layers, W_e is the token embedding matrix, and W_p is the position embedding matrix.

GPT

- GPT 系列的参数量以惊人的速度在不断增长



GPT

- 早期的 GPT 模型（1和2）并没有显著比 BERT 更好，事实上在很多地方不如 BERT
- 但令人惊讶的是，GPT 的“大力出奇迹”模式终于在 GPT-3 上大放异彩

GPT

- 后续 OpenAI 推出的 ChatGPT 更是将公众对该系列模型的关注度提升到了顶峰
- 不过与之相关的技术细节也公开得越来越少
- 有猜测说 ChatGPT 在 GPT-3 的基础上融合了有监督微调、强化学习、思维链等技术
- GPT-4 被认为可能使用了专家模型

扩展阅读

- <https://www.bilibili.com/video/BV1AF411b7xQ/>