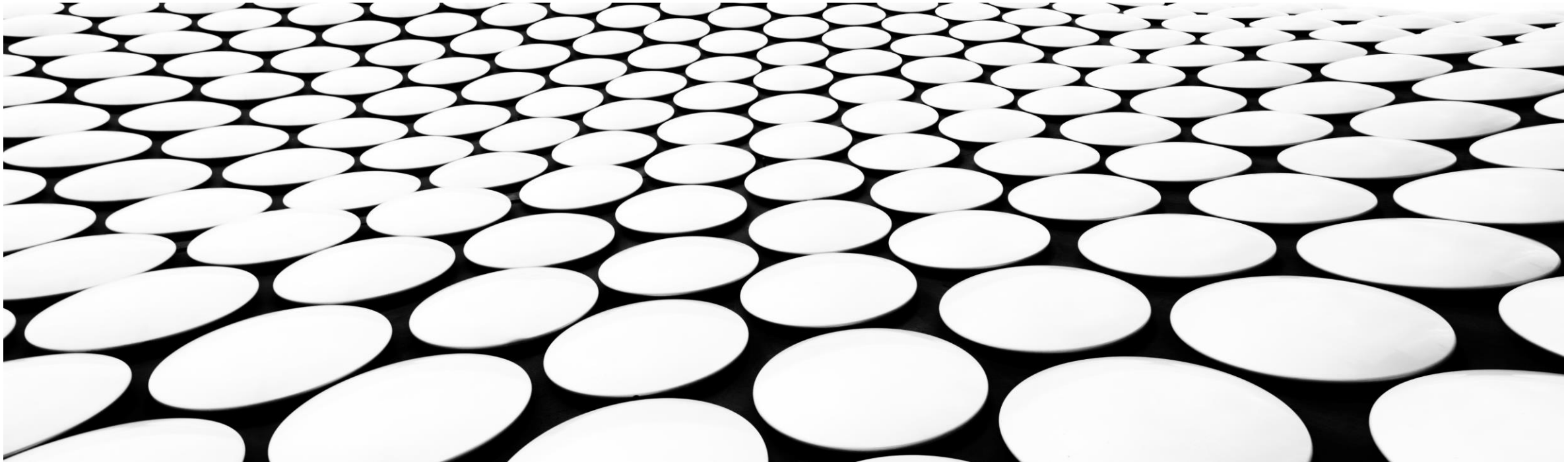
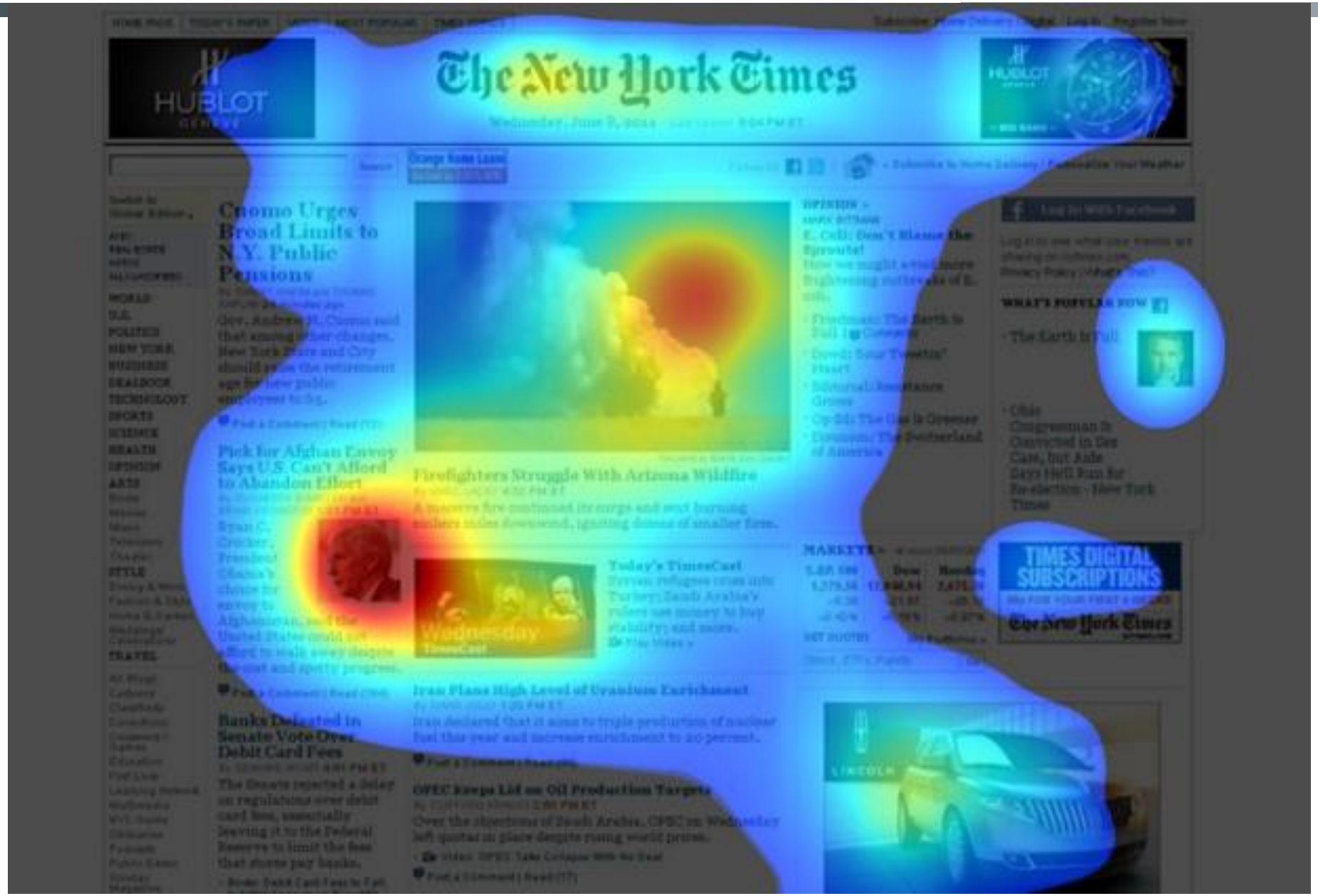
深度学习

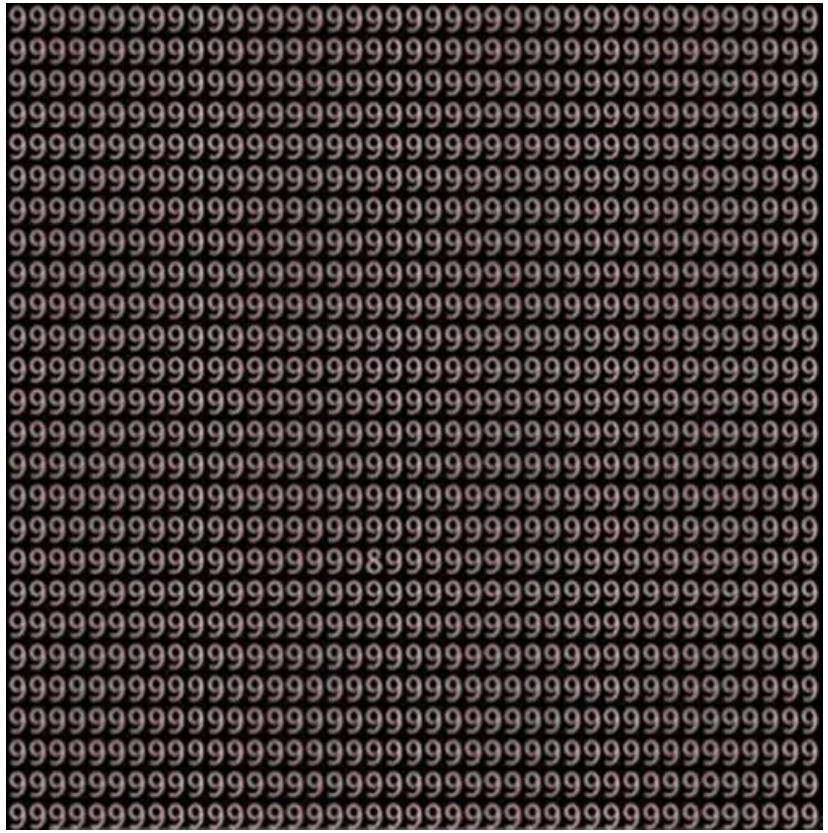
邱怡轩

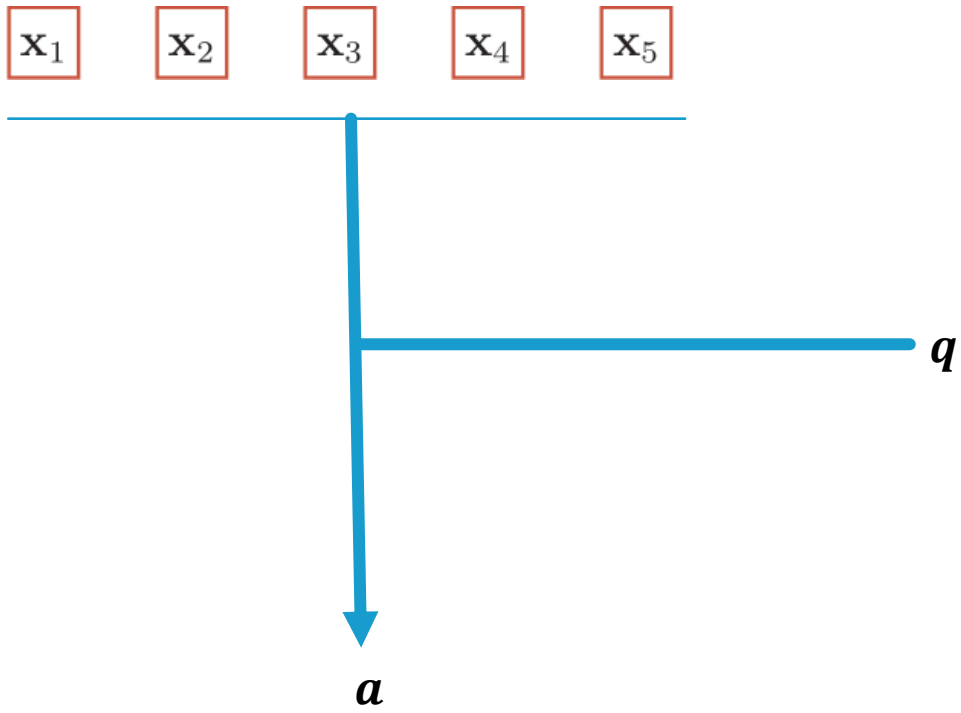


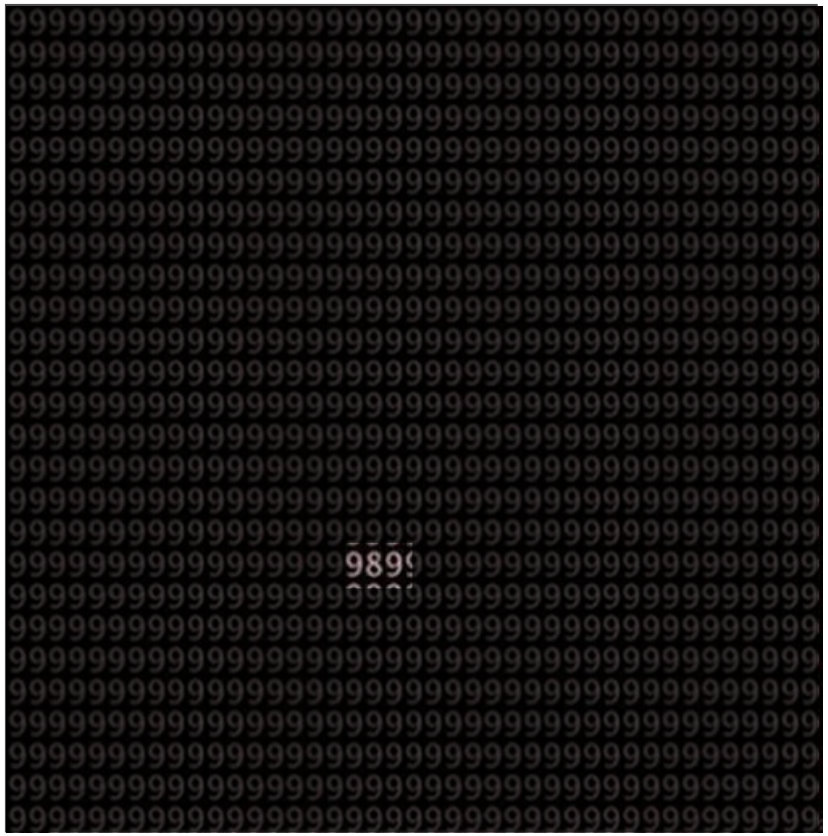
今天的主题

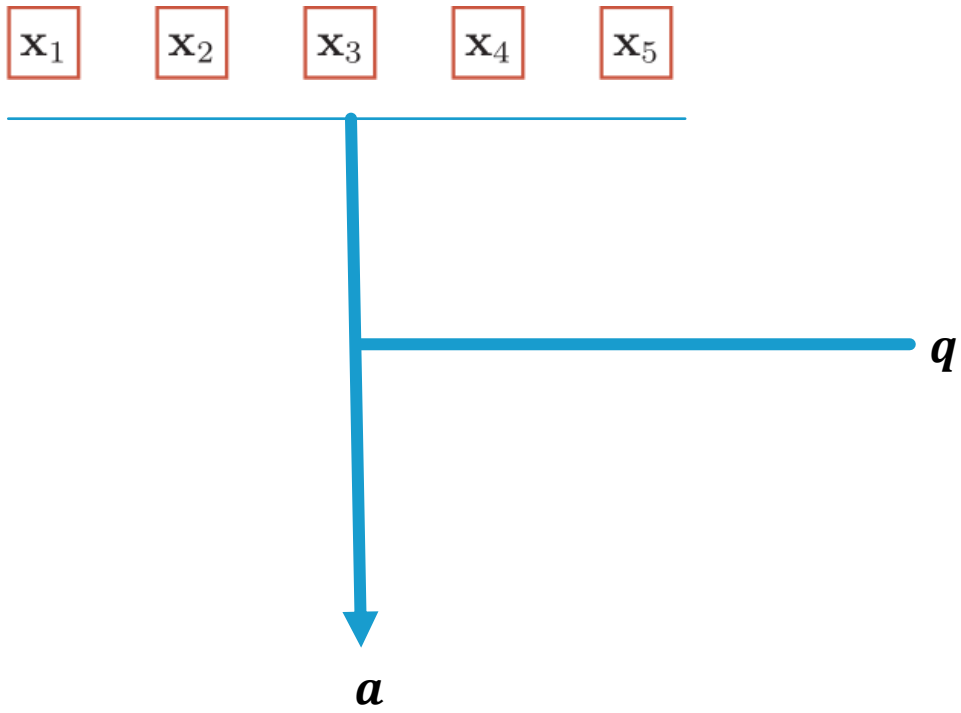
- 注意力机制与 Transformer
- 大语言模型初探


The New York Times
HUBLOT
Cuomo Urges Broad Limits to N.Y. Public Pensions
Pick for Afghan Envoy Says U.S. Can't Afford to Abandon Effort
Banks Defeated in Senate Vote Over Debit Card Fees
Firefighters Struggle With Arizona Wildfire
Wednesday
Iran Plans High Level of Uranium Enrichment
OPEC Keeps Lid on Oil Production Targets
TIMES DIGITAL SUBSCRIPTIONS
LINCOLN



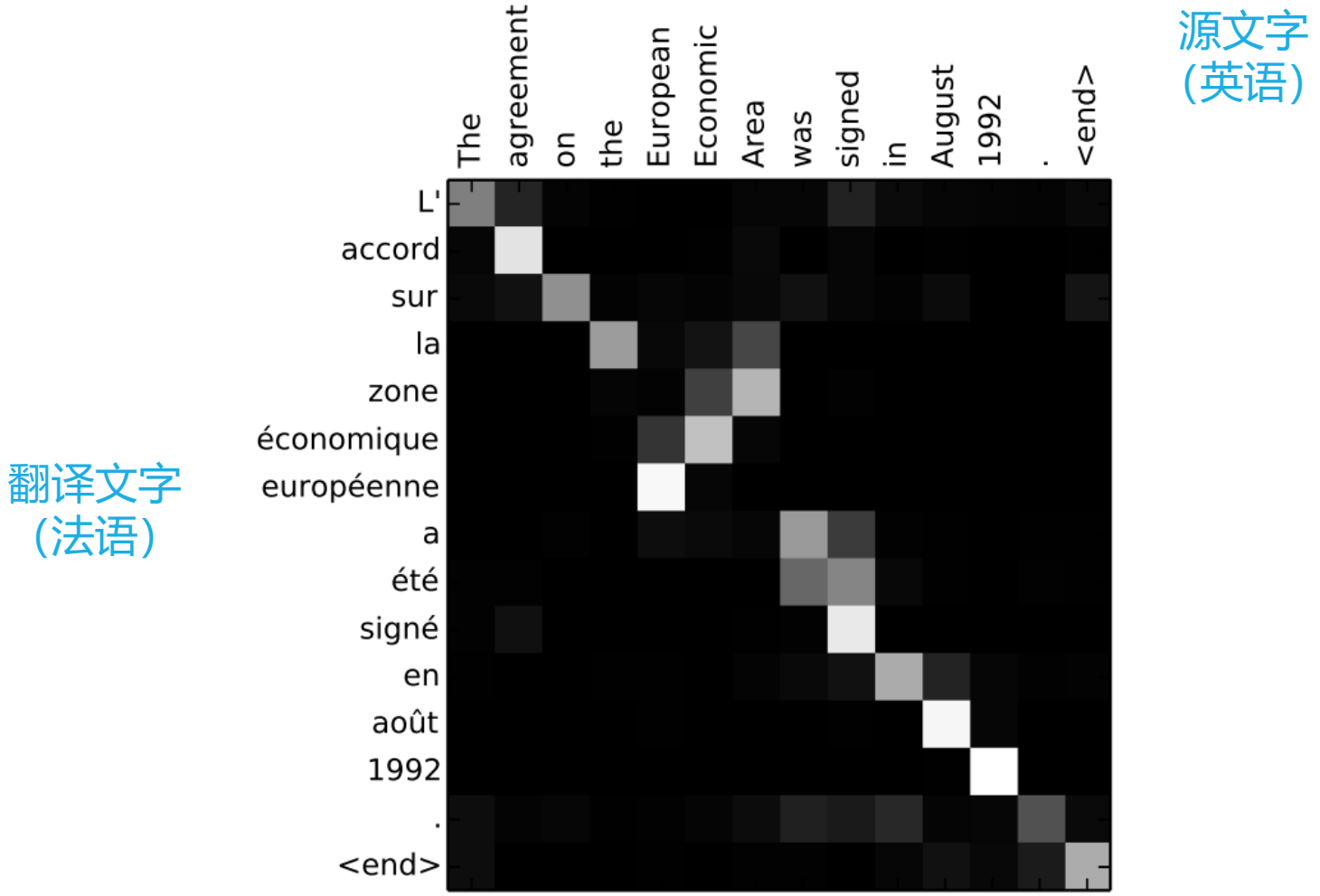

x_1
x_2
x_3
x_4
x_5
q
a




x_1
x_2
x_3
x_4
x_5
q
a

注意力机制

- 在机器翻译中，注意力机制可以理解为一种文字“对齐”的方法



实现方法

- 基于 *Neural Machine Translation by Jointly Learning to Align and Translate*. ICLR 2015.
- 回顾翻译文字的生成机制
- $p(y_i|y_1, \dots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i)$
- y_i 是当前翻译出的单词
- s_i 是 RNN 当前的隐藏层向量，$s_i = f(s_{i-1}, y_{i-1}, c_i)$
- 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

ICLR
International Conference On Learning Representations

APRIL
14
2025

Announcing the Test of Time Award Winners from ICLR 2015

CARL VONDRICK / ICLR 2025

We are honored to announce the Test of Time awards for ICLR 2025. This award recognizes papers published ten years ago at ICLR 2015 that have had a lasting impact on the field. The 2025 program chairs and general chair reviewed the papers published at ICLR 2015, and selected the two papers below for their profound influence and impact on machine learning today.

Congratulations to the authors of the Test of Time winner and runner up!

Runner Up

Neural Machine Translation by Jointly Learning to Align and Translate
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio
https://arxiv.org/abs/1409.0473

Introducing a form of attention, this paper fundamentally changed how sequence-to-sequence models process information. Before this work, encoder-decoder architectures usually compressed entire input sequences into fixed-length vectors, creating memory bottlenecks for longer sequences. The proposed approach enabled the model to “attend” to different parts of the source sentence dynamically during translation, allowing for processing of relevant contextual information. This attention mechanism has since become a cornerstone of modern deep learning, extending far beyond machine translation to form the foundation for transformers and large language models. The paper’s practical impact has been immense, making it one of the most influential contributions to neural network architectures.

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”
- $(\alpha_{i1}, \ldots, \alpha_{iT}) = \mathrm{softmax}(e_{i1}, \ldots, e_{iT})$
- $e_{ij} = a(s_{i-1}, h_j)$ 是一个打分函数，参数数量固定

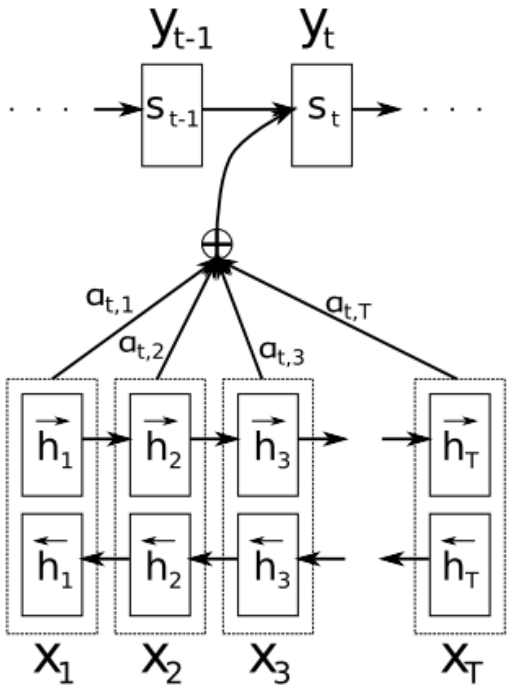


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

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

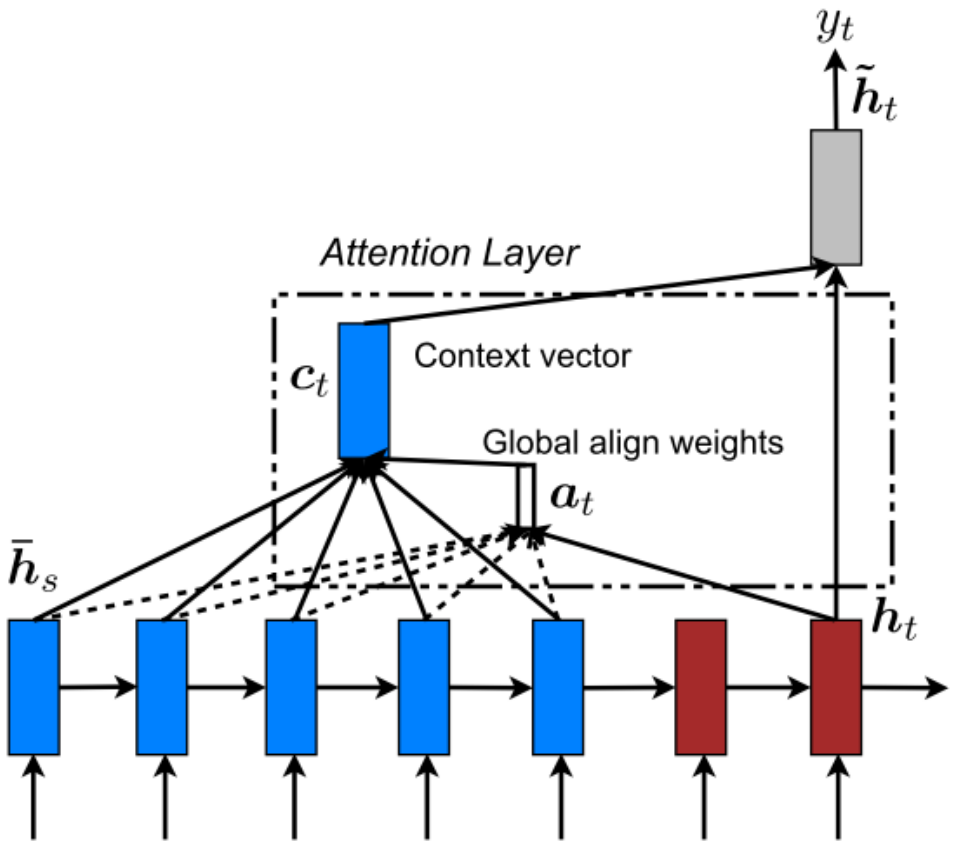
点积模型 $$s(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{x}^\top \boldsymbol{q},$$

缩放点积模型 $$s(\boldsymbol{x}, \boldsymbol{q}) = \frac{\boldsymbol{x}^\top \boldsymbol{q}}{\sqrt{D}},$$

双线性模型 $$s(\boldsymbol{x}, \boldsymbol{q}) = \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{q},$$

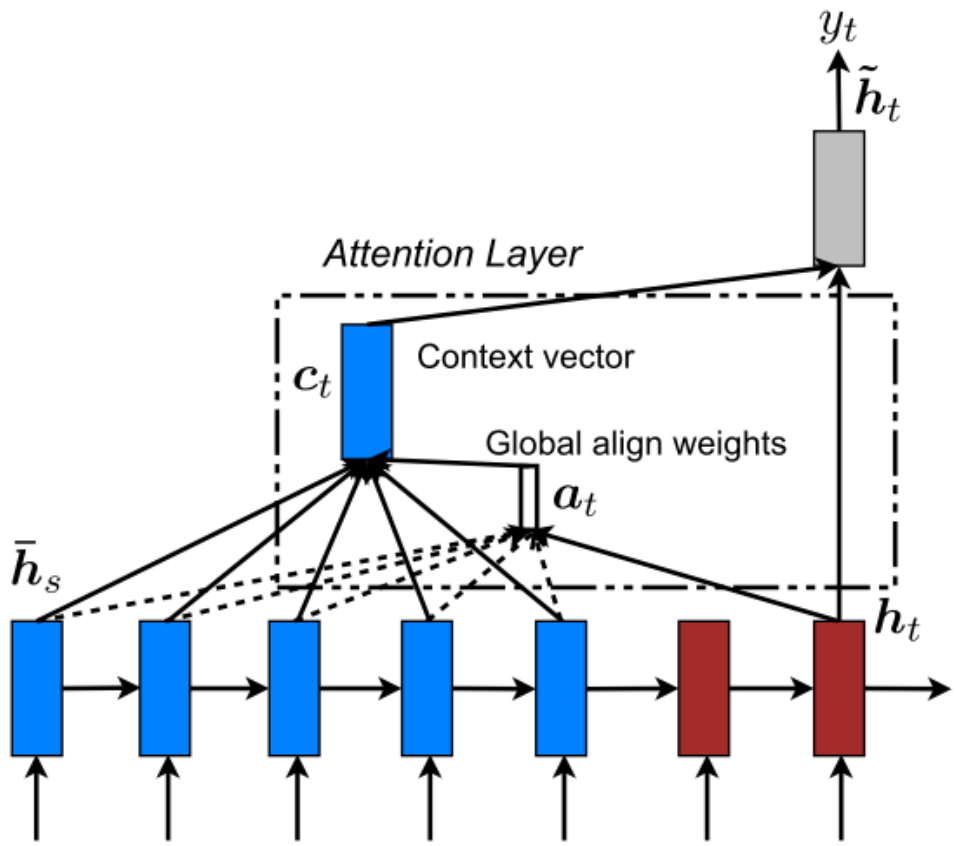
全局注意力

- *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



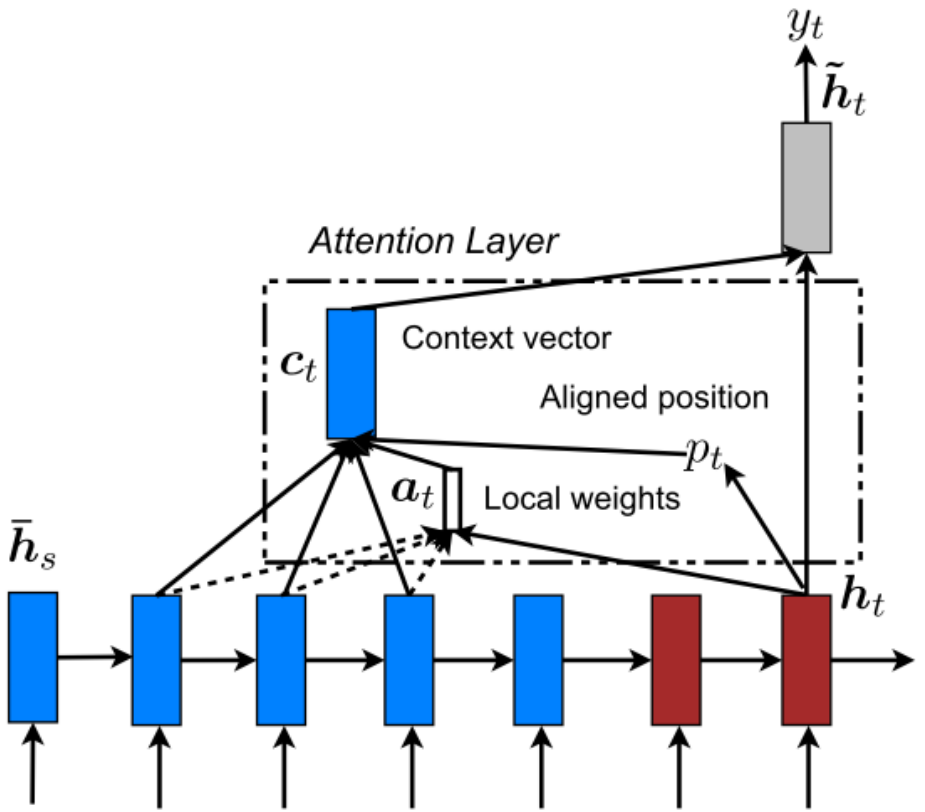
全局注意力

- “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



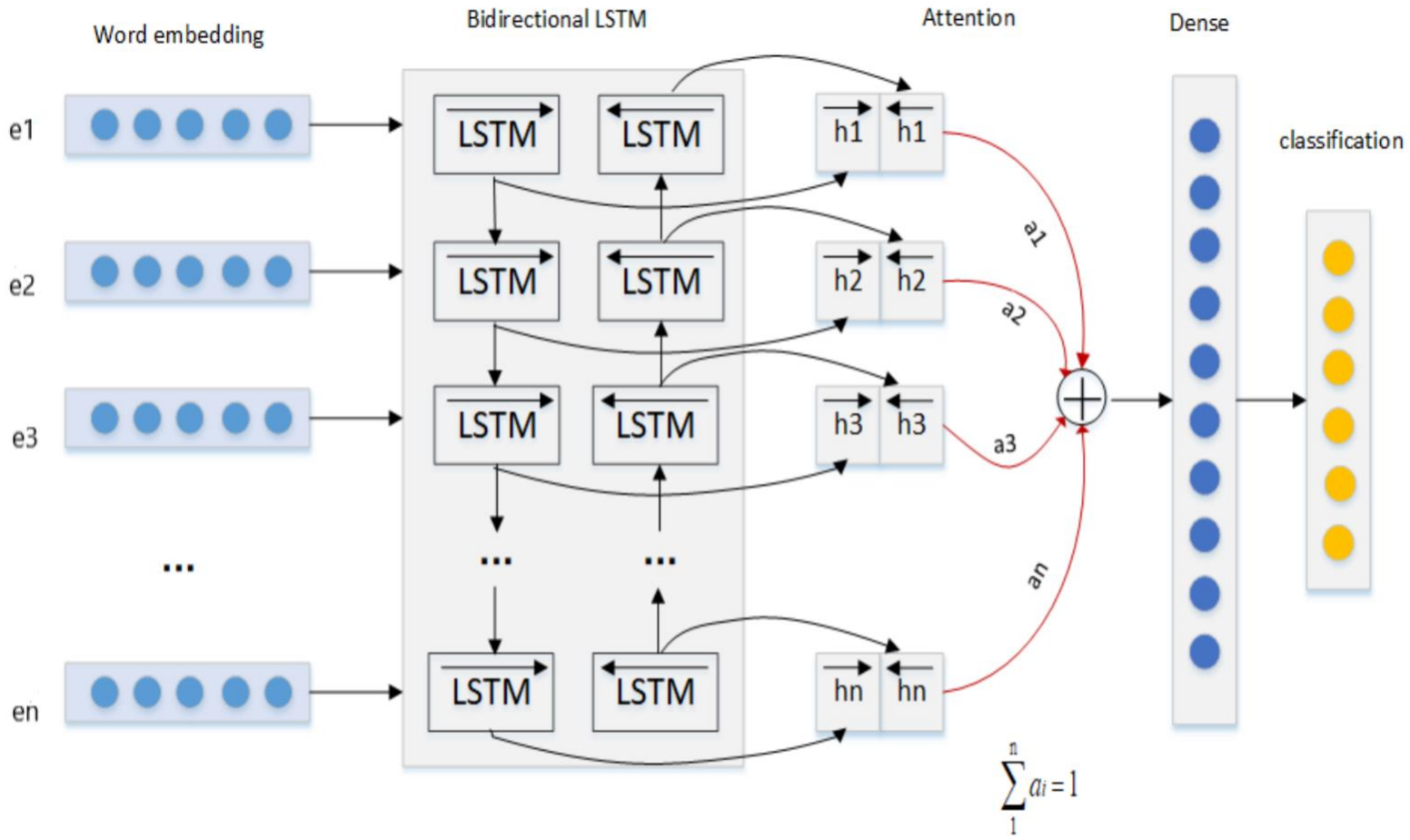
局部注意力

- 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


Word embedding
Bidirectional LSTM
Attention
Dense
classification
e1
e2
e3
...
en
LSTM
h1
h2
h3
hn
a1
a2
a3
an
$\sum_{1}^{n} a_i = 1$

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 (?)

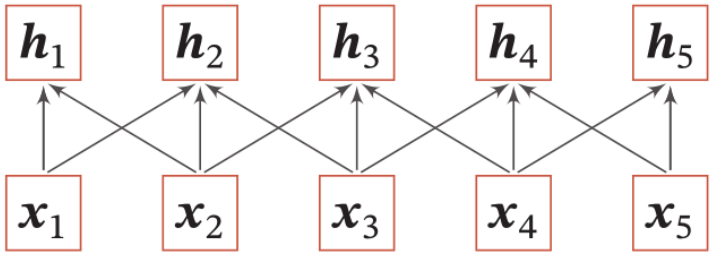
Attention 热潮

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

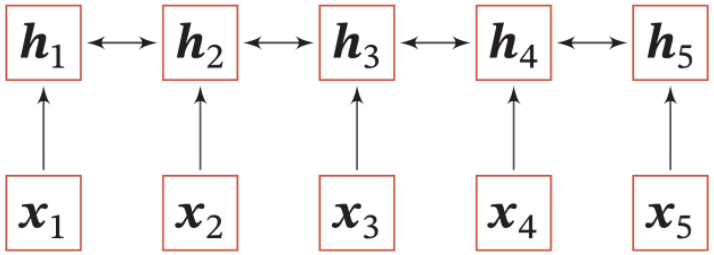
自注意力

- 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



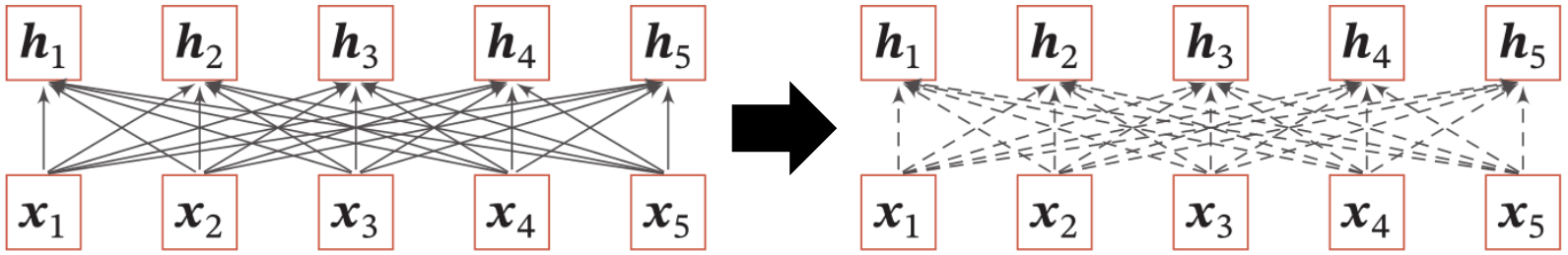
(a) 卷积网络



(b) 双向循环网络

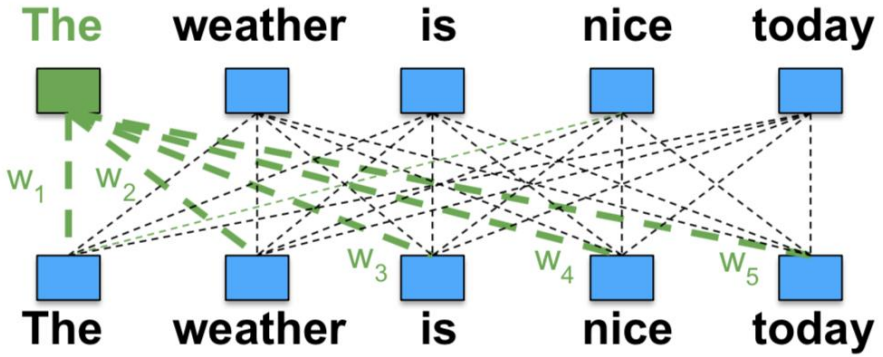
自注意力

- 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



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

$$\begin{bmatrix}1.8\\2.3\\0.4\end{bmatrix}_{\text{The}} = w_1 \times \begin{bmatrix}0.6\\0.2\\0.8\end{bmatrix}_{\text{The}} + w_2 \times \begin{bmatrix}0.2\\0.3\\0.1\end{bmatrix}_{\text{weather}} + w_3 \times \begin{bmatrix}0.9\\0.1\\0.8\end{bmatrix}_{\text{is}} + w_4 \times \begin{bmatrix}0.4\\0.1\\0.4\end{bmatrix}_{\text{nice}} + w_5 \times \begin{bmatrix}0.4\\0.1\\0.6\end{bmatrix}_{\text{today}}$$

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

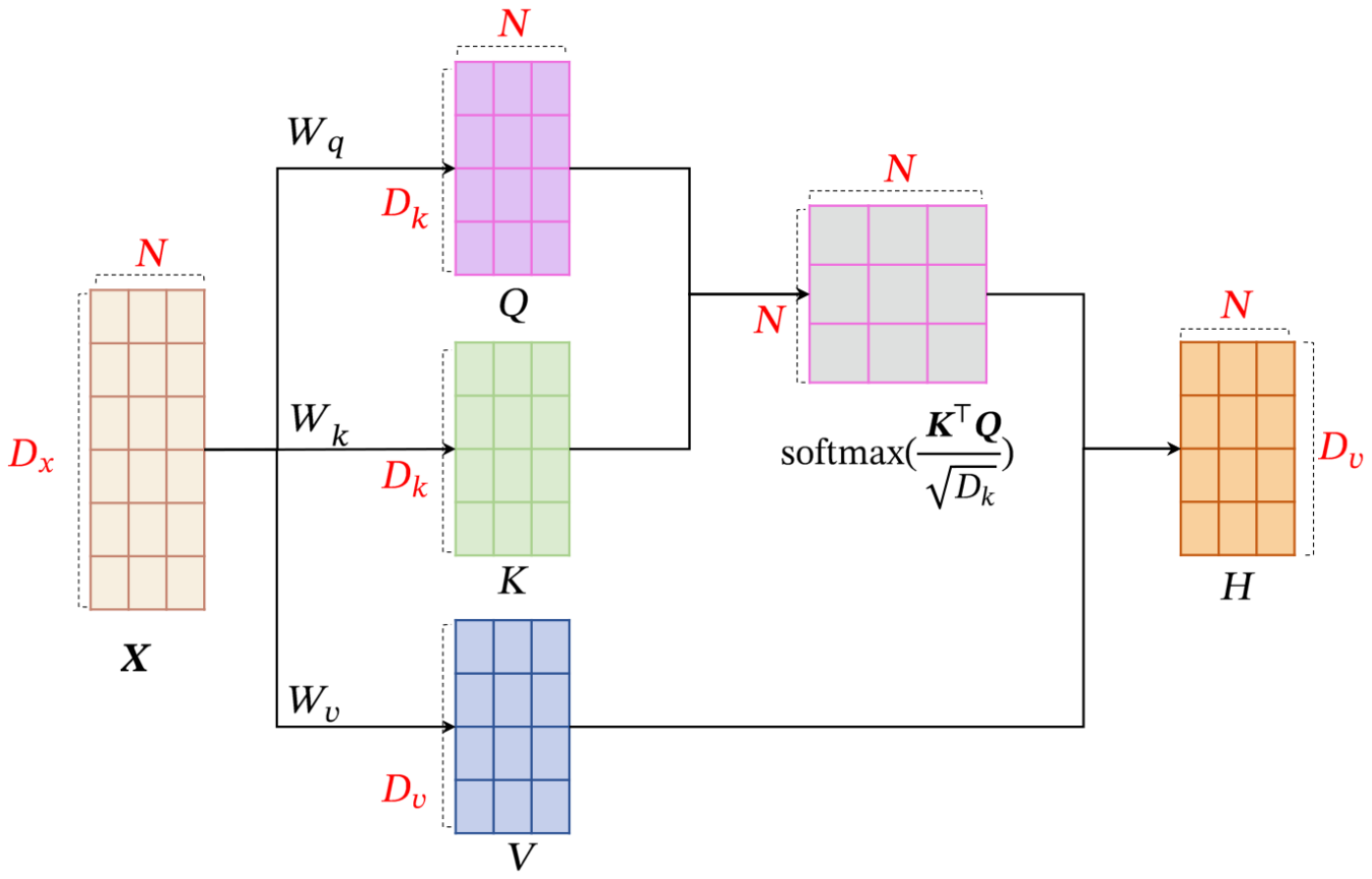
- 上下文向量 $c_i = \sum_{j=1}^{T} \alpha_{ij} h_j$
- 权重 $(\alpha_{i1}, \dots, \alpha_{iT}) = \mathrm{softmax}(e_{i1}, \dots, e_{iT})$
- 打分函数 $e_{ij} = a(s_{i-1}, h_j)$

将 Query 去和 Key 进行匹配

KQV 模式

$$\boldsymbol{Q} = \boldsymbol{W}_q\boldsymbol{X} \in \mathbb{R}^{D_k \times N},$$
$$\boldsymbol{K} = \boldsymbol{W}_k\boldsymbol{X} \in \mathbb{R}^{D_k \times N},$$
$$\boldsymbol{V} = \boldsymbol{W}_v\boldsymbol{X} \in \mathbb{R}^{D_v \times N},$$

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

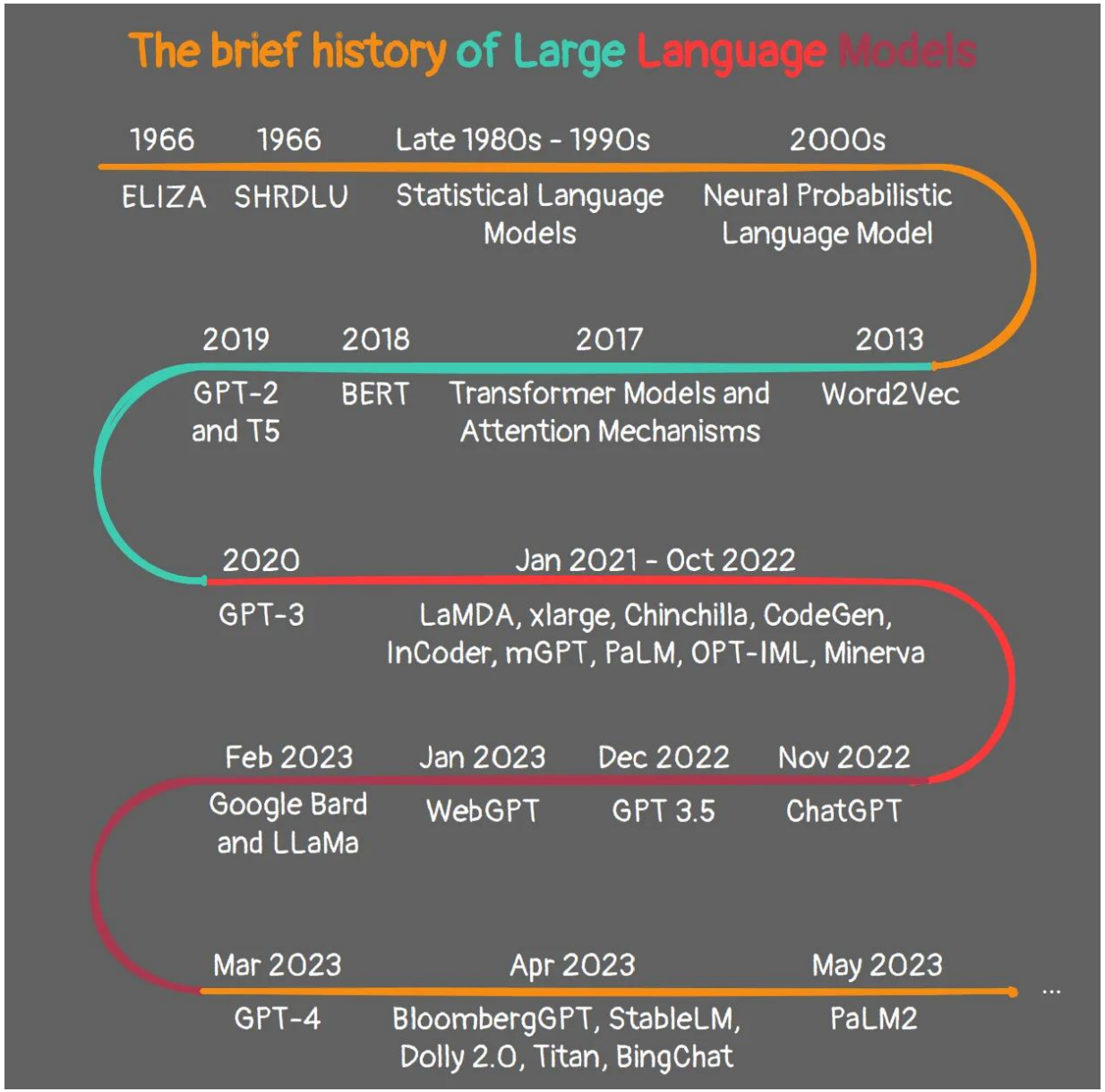


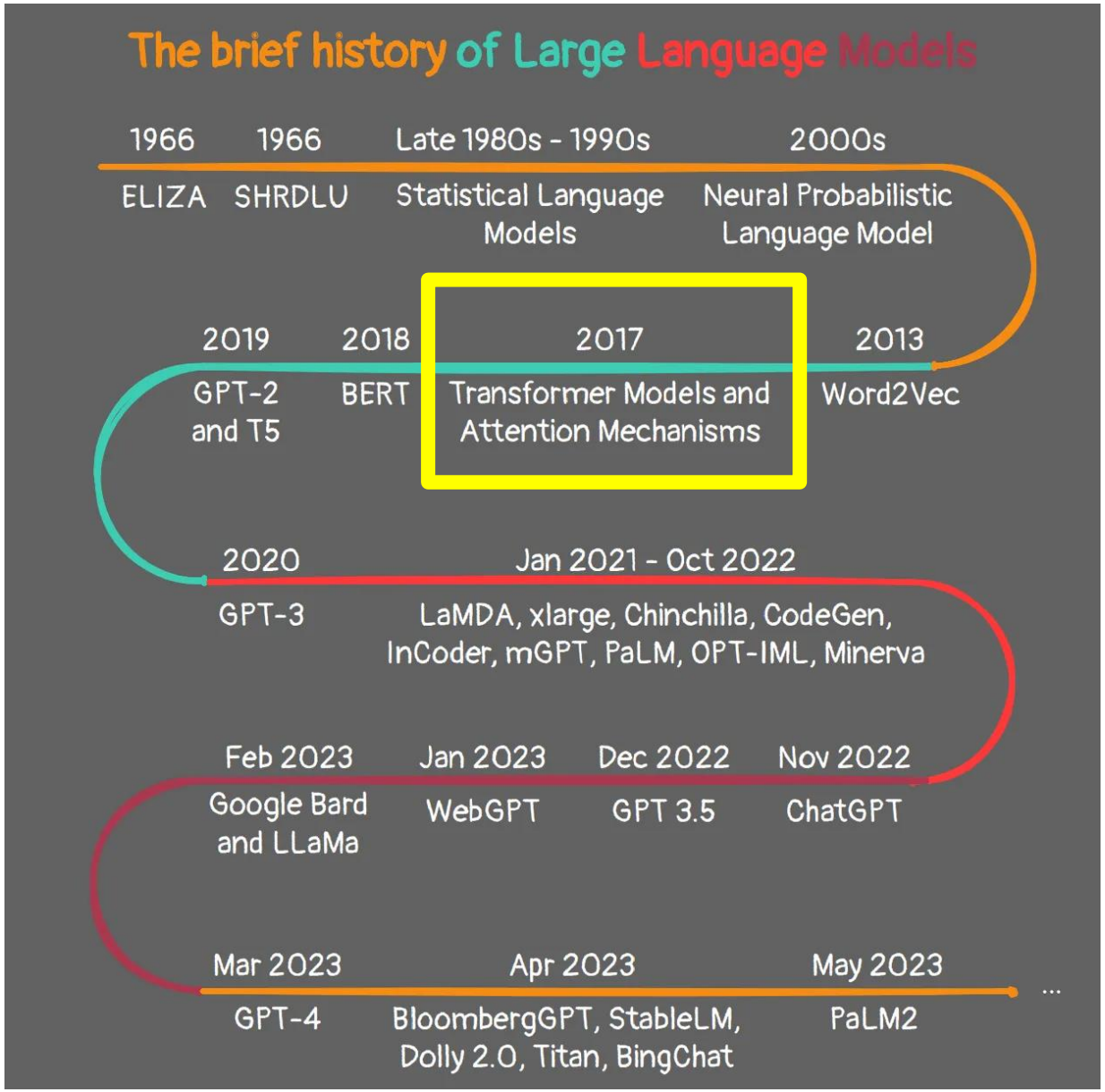
思考

- 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
1966
ELIZA
1966
SHRDLU
Late 1980s - 1990s
Statistical Language Models
2000s
Neural Probabilistic Language Model
2013
Word2Vec
2017
Transformer Models and Attention Mechanisms
2018
BERT
2019
GPT-2 and T5
2020
GPT-3
Jan 2021 - Oct 2022
LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva
Nov 2022
ChatGPT
Dec 2022
GPT 3.5
Jan 2023
WebGPT
Feb 2023
Google Bard and LLaMa
Mar 2023
GPT-4
Apr 2023
BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat
May 2023
PaLM2
...


The brief history of Large Language Models
1966
ELIZA
1966
SHRDLU
Late 1980s - 1990s
Statistical Language Models
2000s
Neural Probabilistic Language Model
2013
Word2Vec
2017
Transformer Models and Attention Mechanisms
2018
BERT
2019
GPT-2 and T5
2020
GPT-3
Jan 2021 - Oct 2022
LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva
Nov 2022
ChatGPT
Dec 2022
GPT 3.5
Jan 2023
WebGPT
Feb 2023
Google Bard and LLaMa
Mar 2023
GPT-4
Apr 2023
BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat
May 2023
PaLM2
...

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

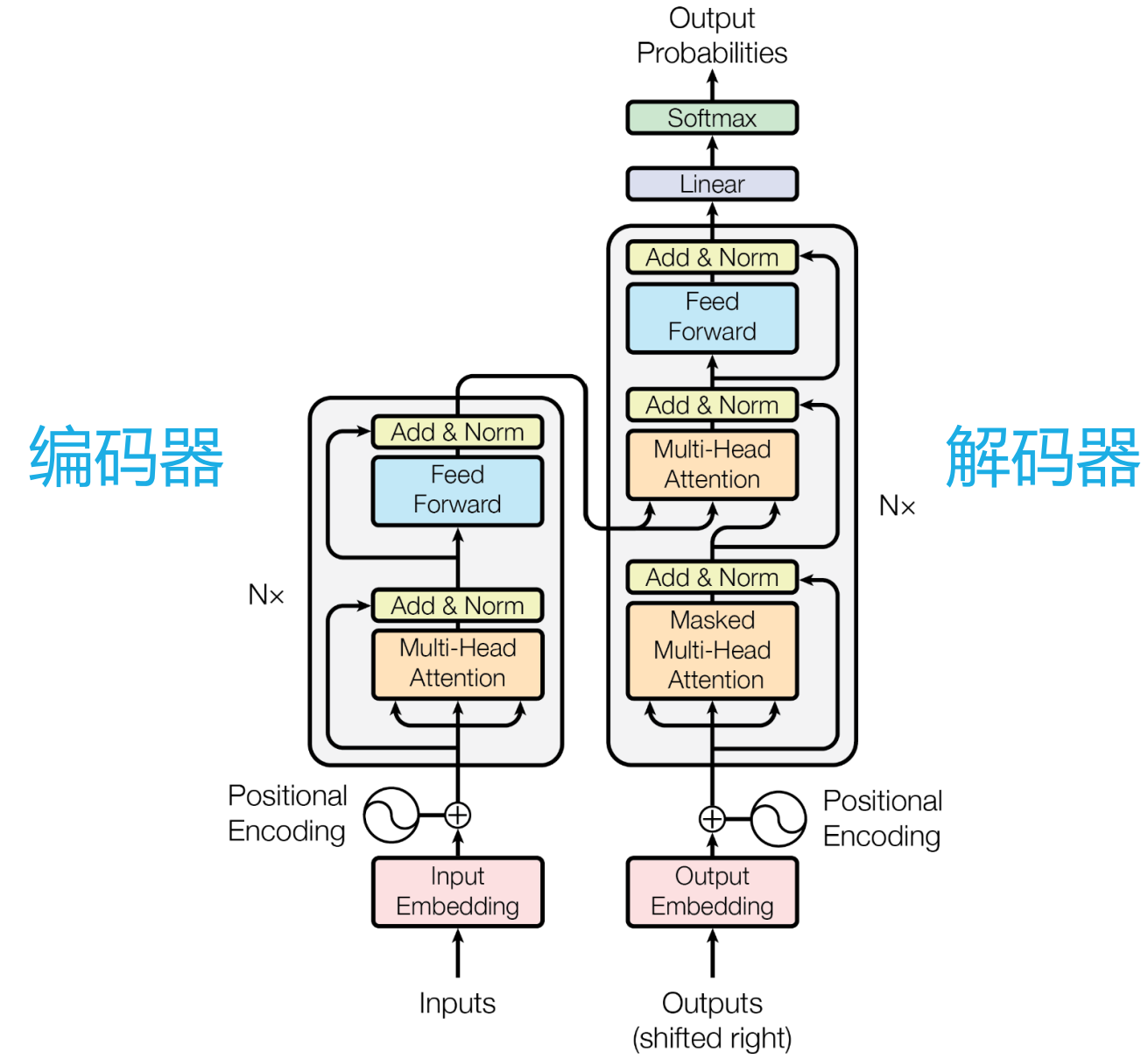
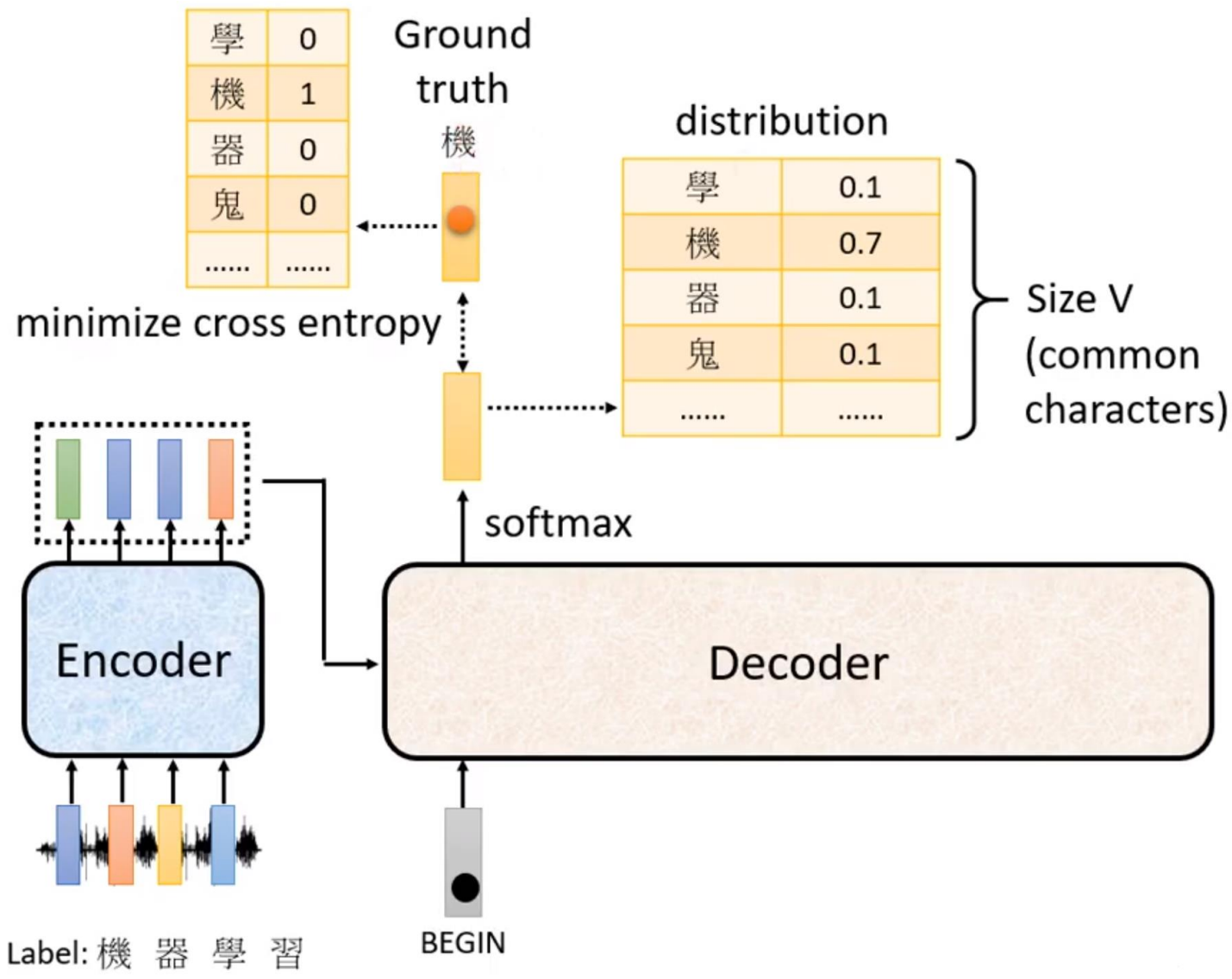


Figure 1: The Transformer - model architecture.

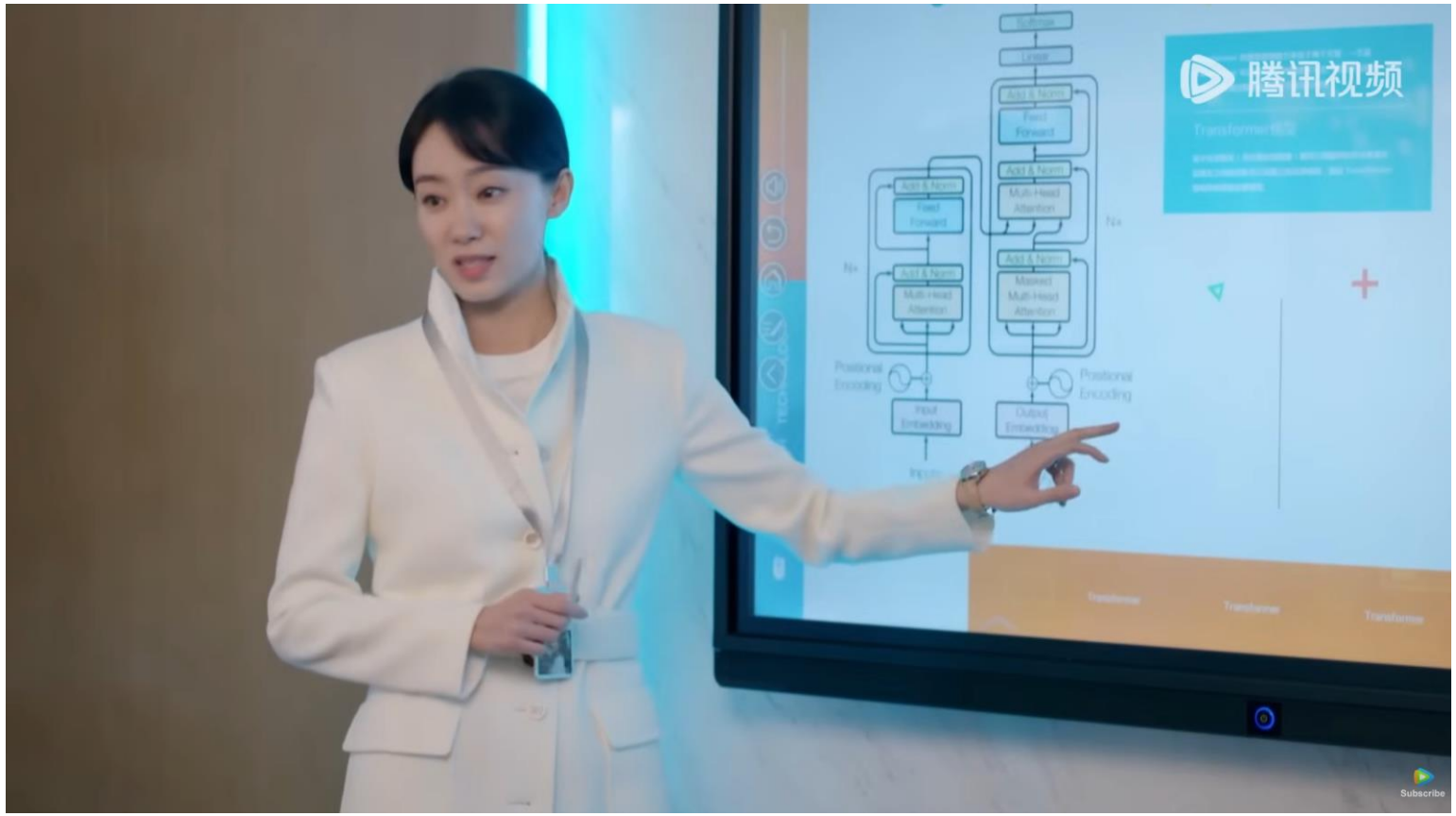
Transformer

本页图片取自李宏毅
Transformer 讲义



Transformer

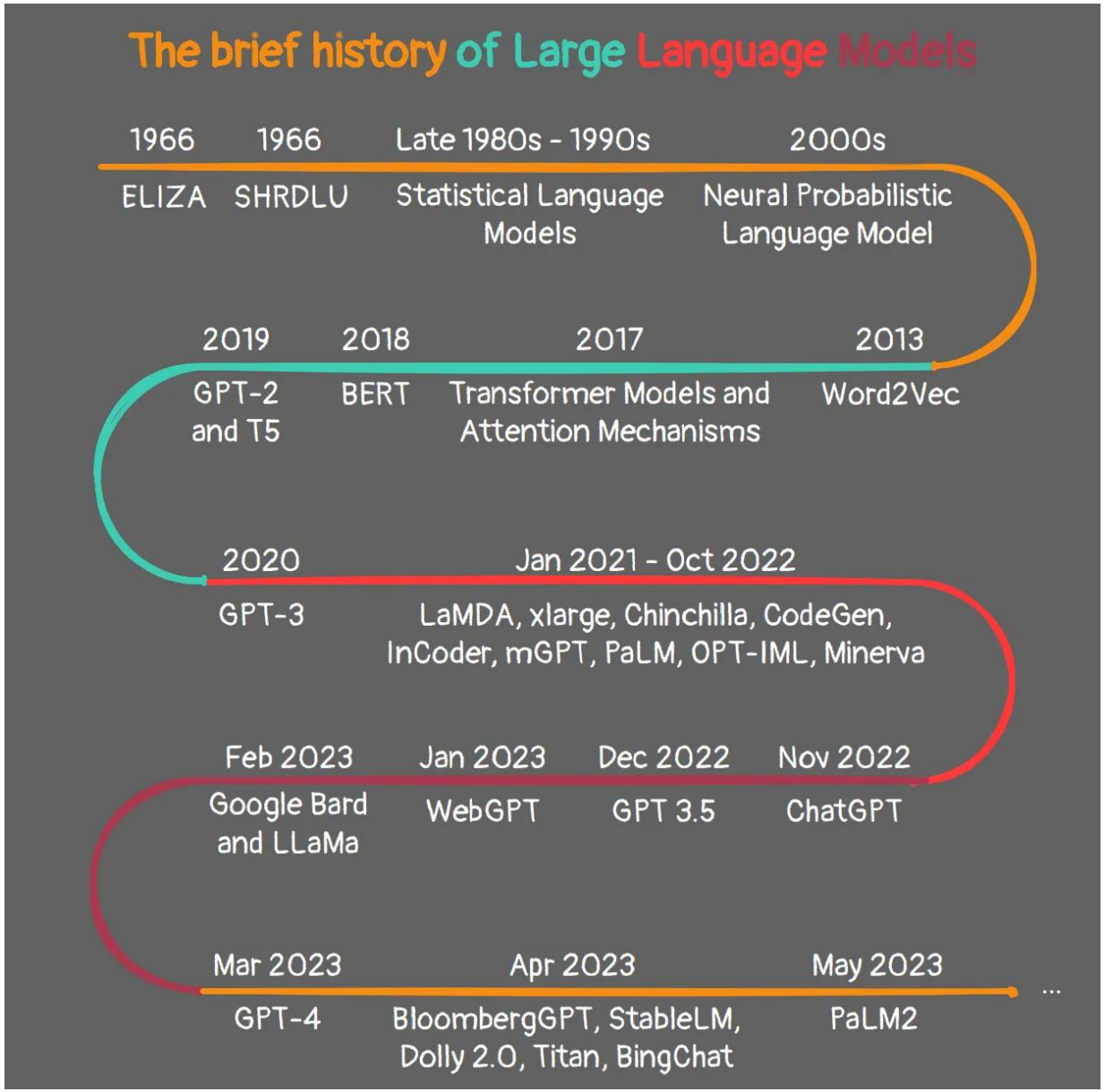
- Transformer 的影响力之大，使其甚至进入了一些流行文化和影视作品之中

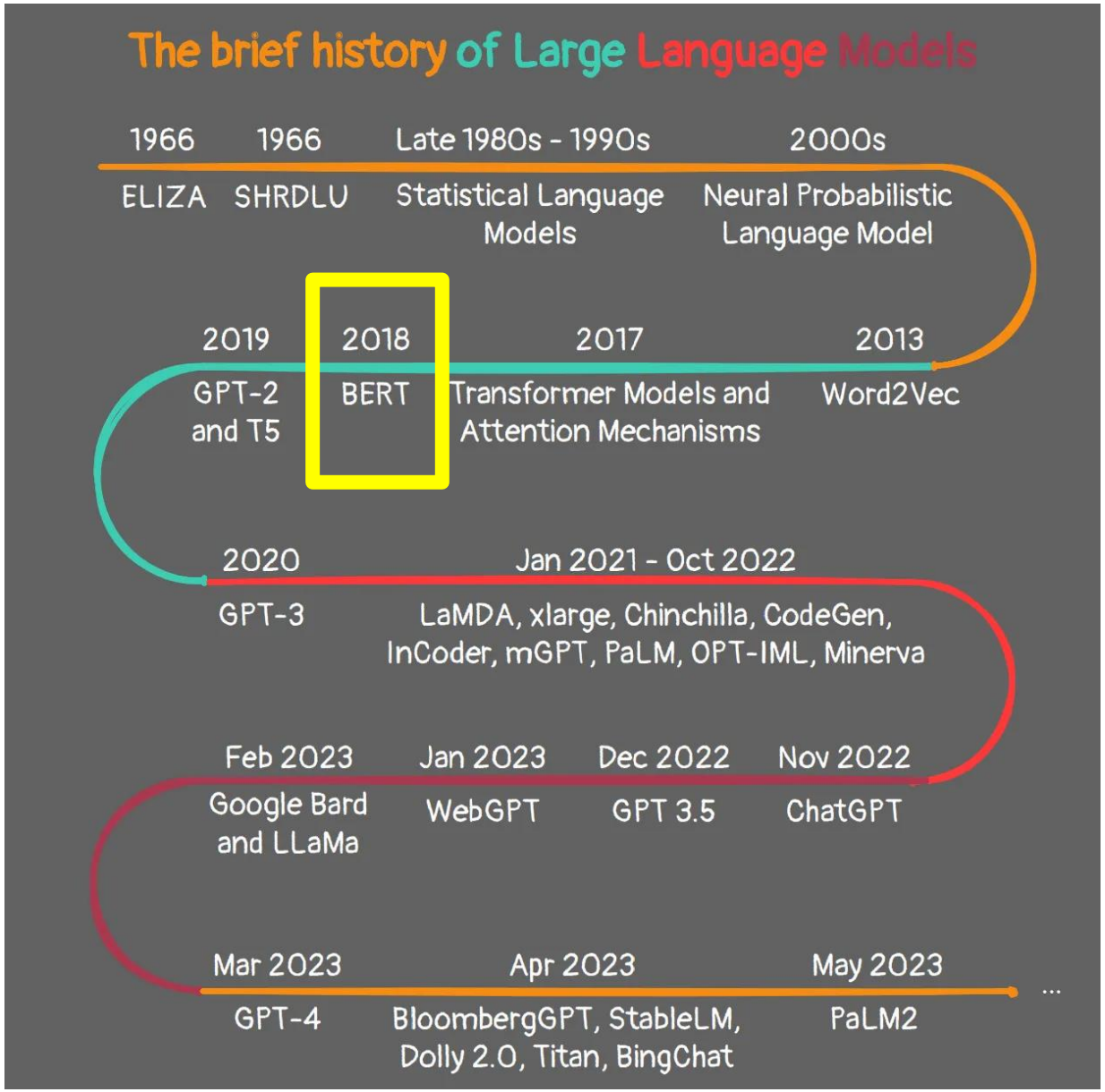


2023年热播电视剧《好事成双》

扩展阅读

- https://www.bilibili.com/video/BV1pu411o7BE
- https://www.bilibili.com/video/BV13z421U7cs
- https://www.bilibili.com/video/BV1TZ421j7Ke


The brief history of Large Language Models
1966
ELIZA
1966
SHRDLU
Late 1980s - 1990s
Statistical Language Models
2000s
Neural Probabilistic Language Model
2013
Word2Vec
2017
Transformer Models and Attention Mechanisms
2018
BERT
2019
GPT-2 and T5
2020
GPT-3
Jan 2021 - Oct 2022
LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva
Nov 2022
ChatGPT
Dec 2022
GPT 3.5
Jan 2023
WebGPT
Feb 2023
Google Bard and LLaMa
Mar 2023
GPT-4
Apr 2023
BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat
May 2023
PaLM2
...


The brief history of Large Language Models
1966
ELIZA
1966
SHRDLU
Late 1980s - 1990s
Statistical Language Models
2000s
Neural Probabilistic Language Model
2013
Word2Vec
2017
Transformer Models and Attention Mechanisms
2018
BERT
2019
GPT-2 and T5
2020
GPT-3
Jan 2021 - Oct 2022
LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva
Nov 2022
ChatGPT
Dec 2022
GPT 3.5
Jan 2023
WebGPT
Feb 2023
Google Bard and LLaMa
Mar 2023
GPT-4
Apr 2023
BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat
May 2023
PaLM2
...

BERT

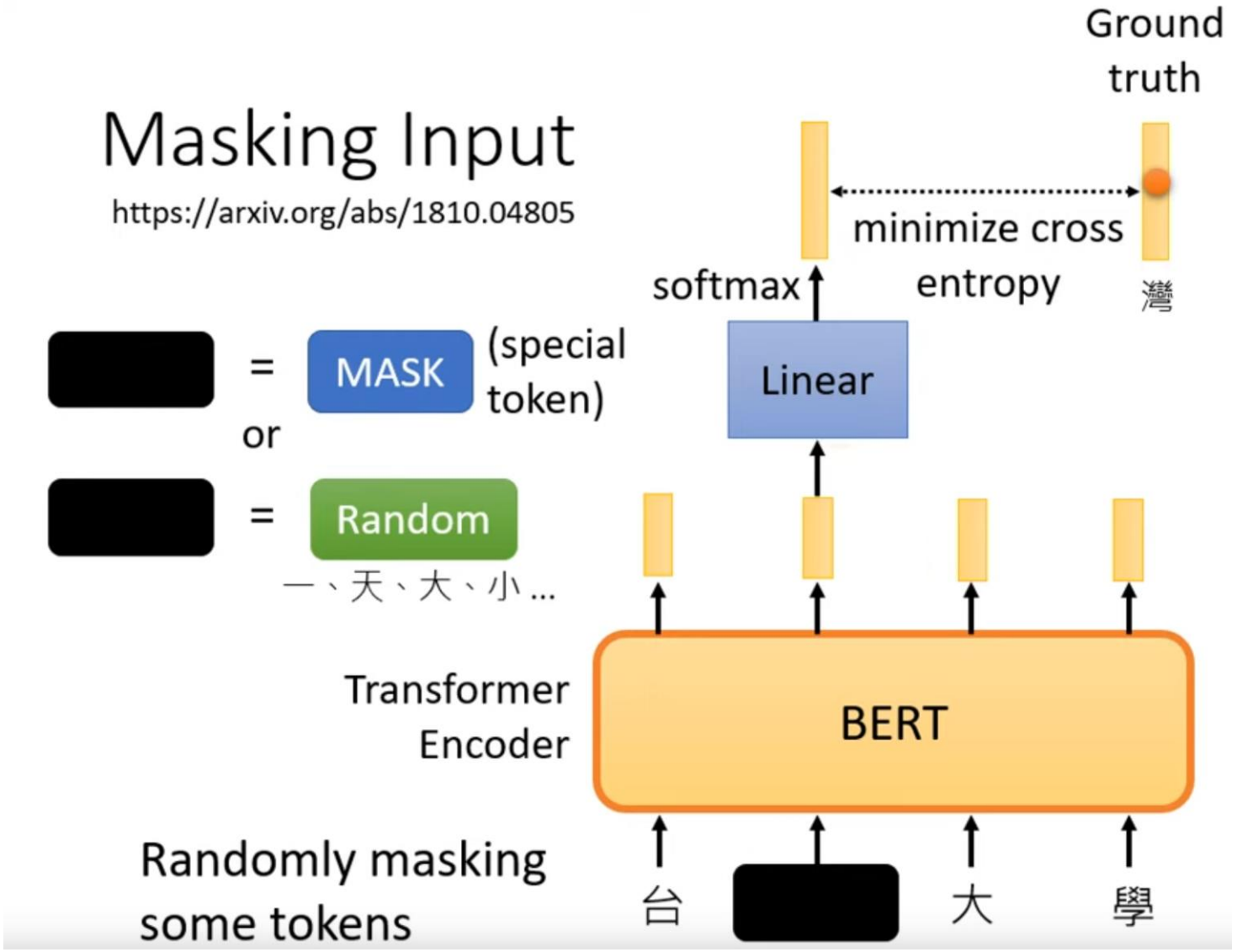
- BERT 的全称为 Bidirectional Encoder Representations from Transformers
- 同样是利用了 Transformer 结构构建出的语言模型

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

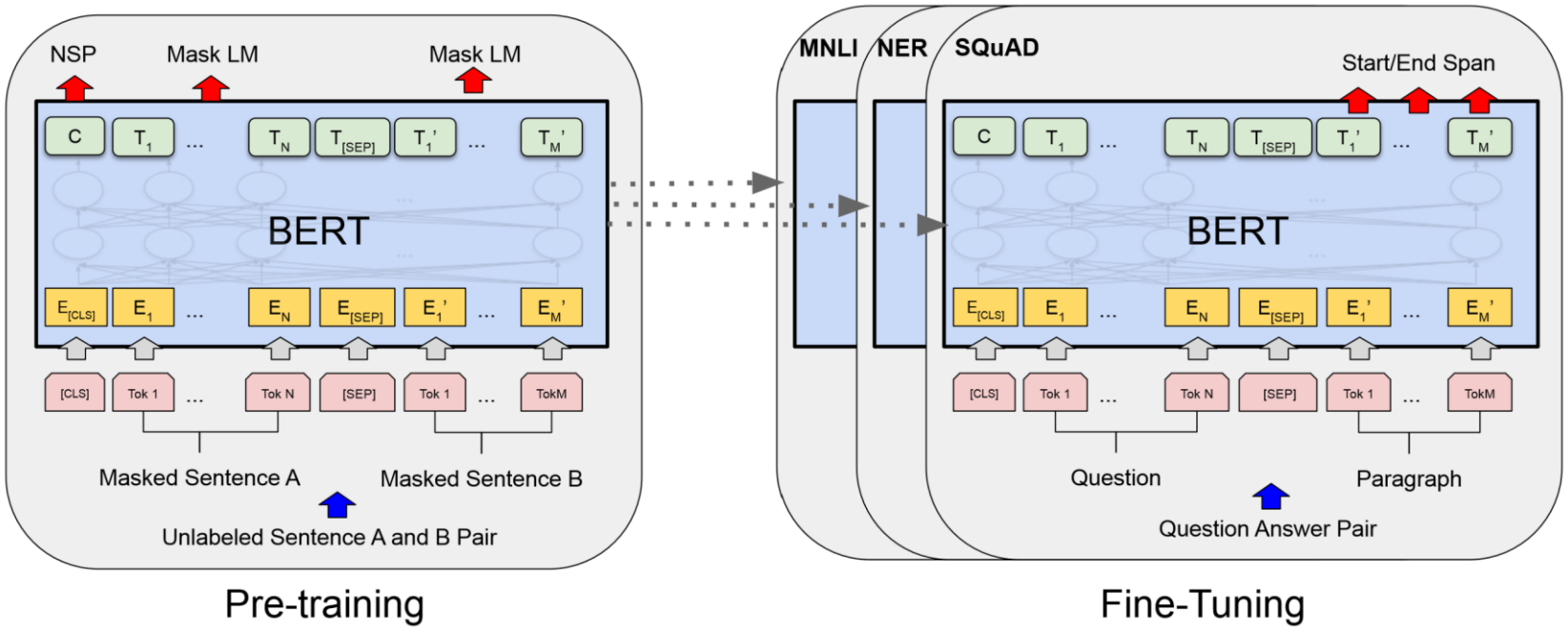
BERT

本页图片取自李宏毅
BERT 讲义



BERT

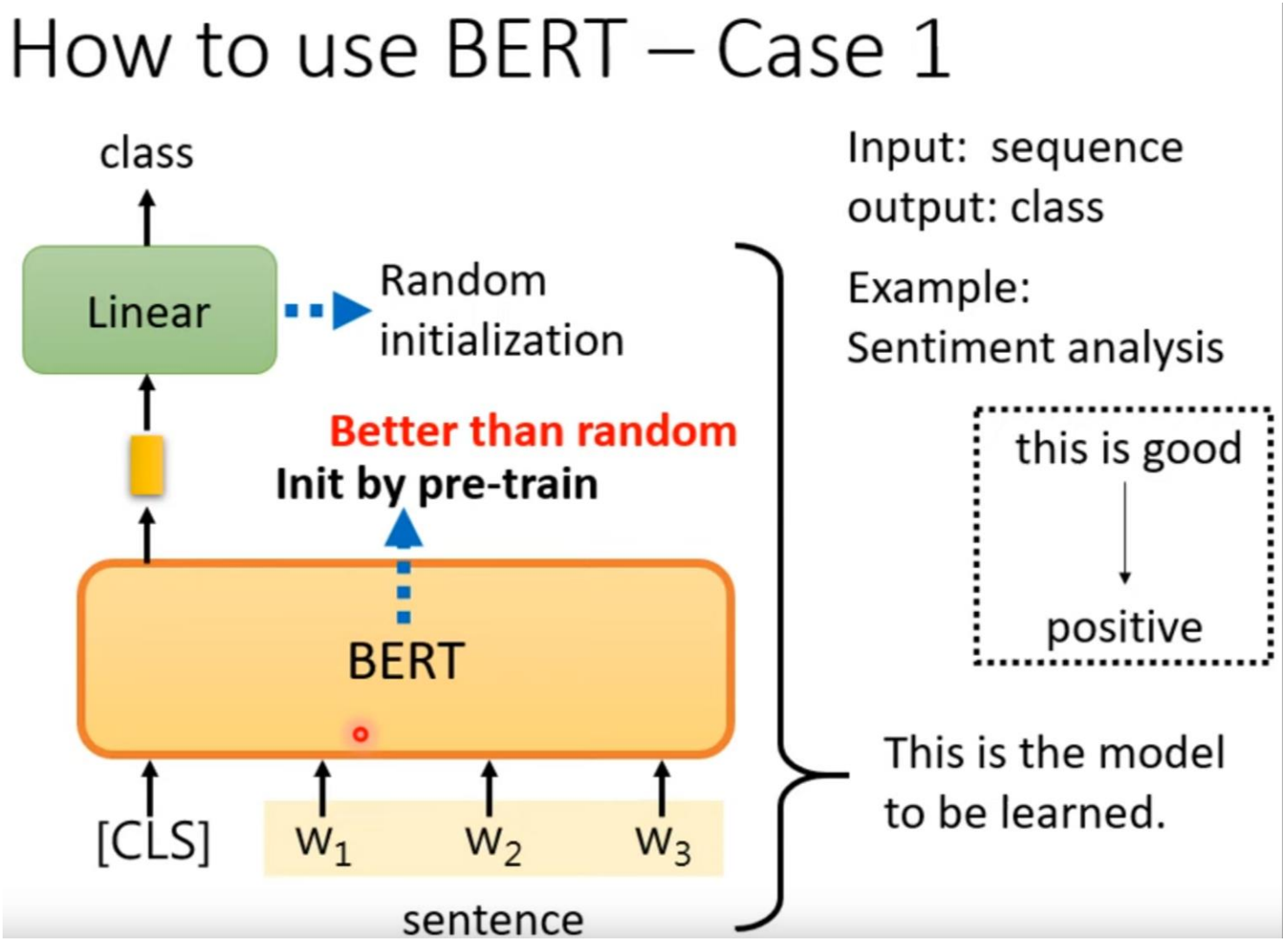
- BERT 极大地普及了“预训练+微调”这一大模型的使用模式
- 在训练阶段获取文本的嵌入向量表达
- 微调阶段用向量表达完成各项 NLP 任务



BERT

本页图片取自李宏毅

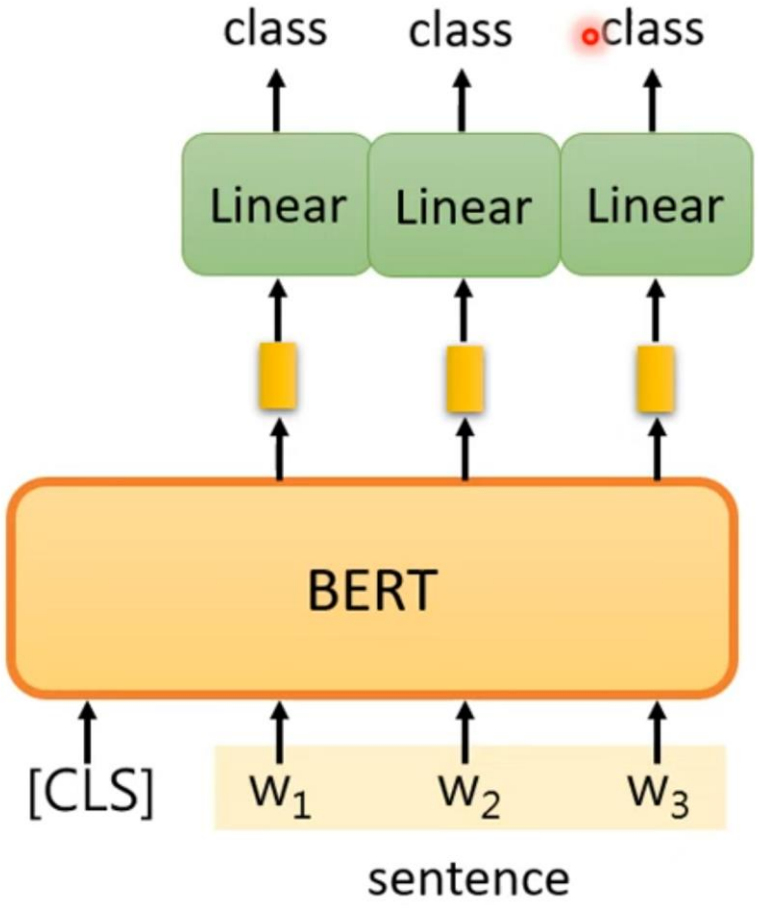
BERT 讲义



BERT

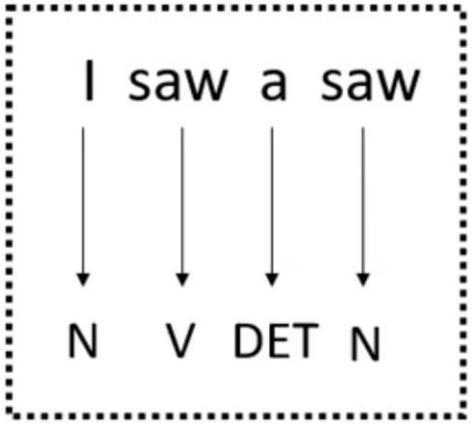
本页图片取自李宏毅
BERT 讲义

How to use BERT – Case 2



Input: sequence
output: same as input

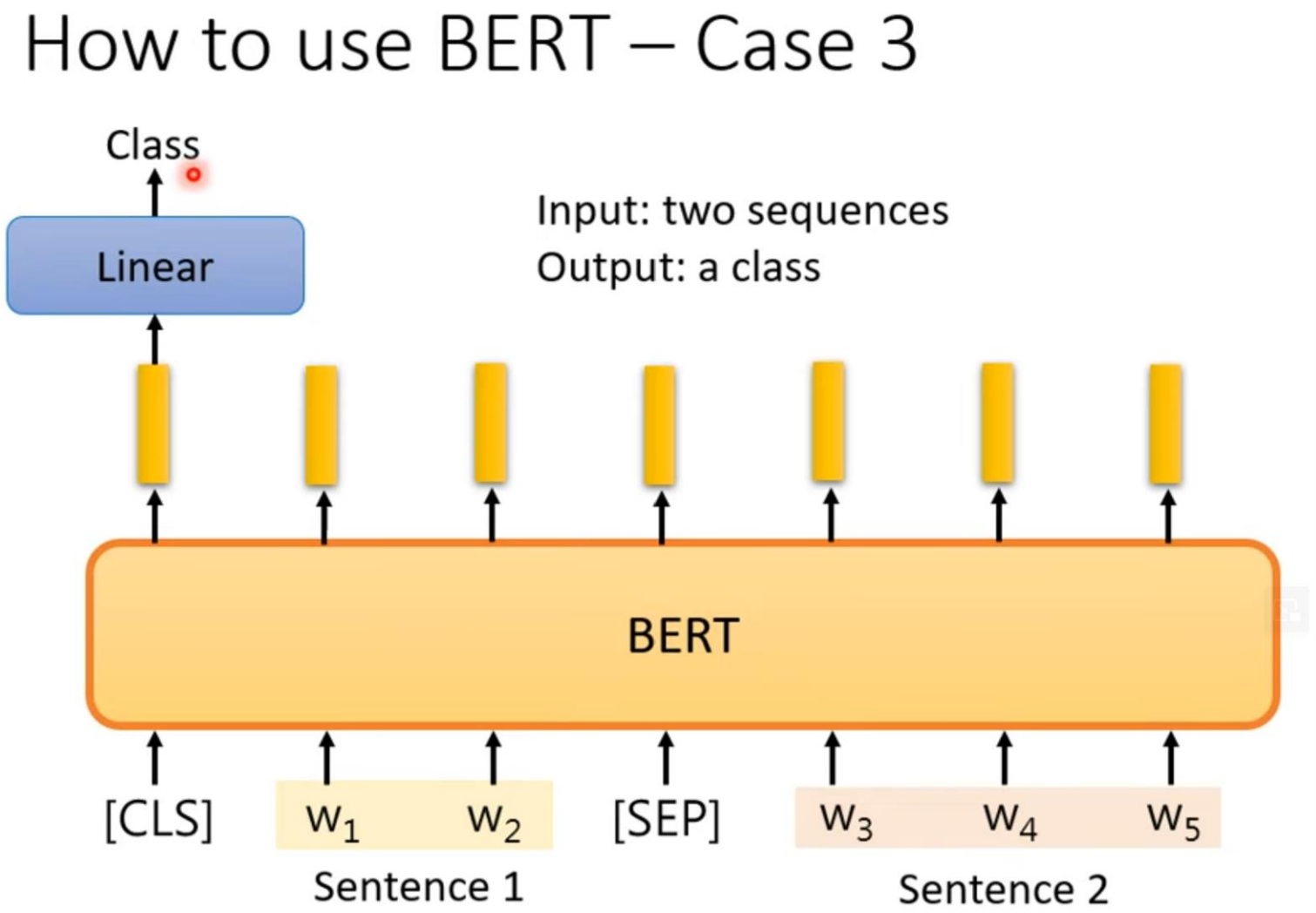
Example:
POS tagging



BERT

本页图片取自李宏毅

BERT 讲义



扩展阅读

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