预训练（预热）

CS2916 大语言模型

饮水思源 爱国荣校

https://plms.ai/teaching/index.html

预训练

- 以**语言模型（类语言模型）**为优化准则在大规模文本语料上进行无监督学习

例 5　如图 6.3-13，已知□$ABCD$ 的三个顶点 A，B，C 的坐标分别是 $(-2, 1)$，$(-1, 3)$，$(3, 4)$，求顶点 D 的坐标.

解法 1：如图 6.3-13，设顶点 D 的坐标为 (x, y).

因为 $\overrightarrow{AB}=(-1-(-2), 3-1)=(1, 2)$，

$\overrightarrow{DC}=(3-x, 4-y)$，

又 $\overrightarrow{AB}=\overrightarrow{DC}$，

所以 $(1, 2)=(3-x, 4-y)$.

即 $\begin{cases}1=3-x,\\2=4-y,\end{cases}$ 解得 $\begin{cases}x=2,\\y=2.\end{cases}$

所以顶点 D 的坐标为 $(2, 2)$.

准确预测出下一个词是一件并不容易的事情

中国的首都是___

准确预测出下一个词是一件并不容易的事情

中国的首都是___

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？___

准确预测出下一个词是一件并不容易的事情

中国的首都是___

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？___

这天，柯南收到了一封来自大版的信…(此处省略数千字)…凶手是___

准确预测出下一个词是一件并不容易的事情

中国的首都是__

小芬对小芳说：“后天的大前天的后天，也就是昨天的昨天的大后天是我的生日，请来参加我的生日会。”小芳应该什么时候赴约呢？__

这天，柯南收到了一封来自大版的信…(此处省略数千字)…凶手是__

1234567 * 54321 + 1234567 / 2 = __

建模世界所有的文本

$$P(w_1, \cdots, w_T)$$

建模世界

$$P(w_1, \cdots, w_T)$$

- Humanoid Locomotion as Next Token Prediction arXiv 2024
- Genie: Generative Interactive Environments, arXiv 2024

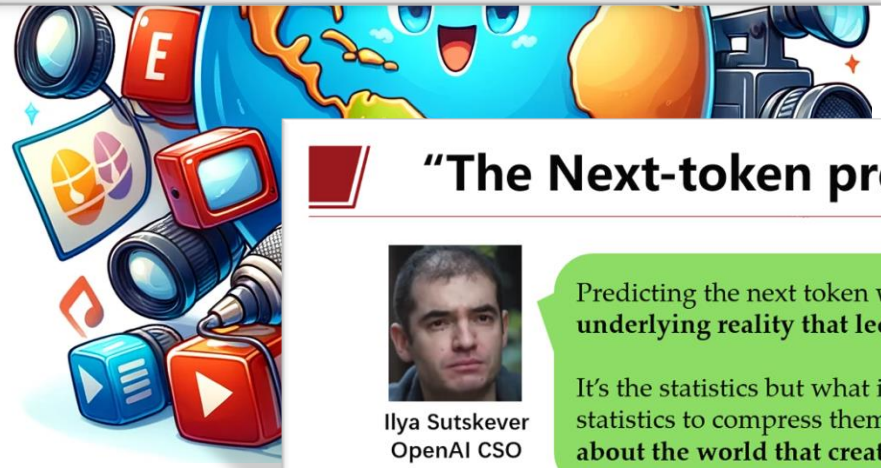
建模世界

"The Bitter Lesson"

Rich Sutton
强化学习之父

The biggest lesson that can be read from 70 years of AI research is that general methods that **leverage computation** are ultimately the most effective, and by a large margin

We want AI agents that can **discover like we can**, not which contain what we have discovered. Building in our discoveries only makes it harder to see how the discovering process can be done.



"The Next-token prediction is enough for AGI"

Ilya Sutskever
OpenAI CSO

Predicting the next token well means that **you understand the underlying reality that led to the creation of that token.**

It's the statistics but what is statistics? In order to understand those statistics to compress them, you need to **understand what is it about the world that creates those statistics**

课程内容

- 语言模型
  - 理解语言模型基本概念和评估方法
  - 了解语言模型常见应用
  - 掌握基于统计、和基于神经网络的学习方法
- 表示学习
  - 了解词表示的学习概念和意义
  - 掌握基于神经网络的词表示学习方法
    - word2vec的基本原理
  - 了解不同词表示学习方法的差异
  - 了解句子表示的学习概念和意义
  - 掌握基于神经网络的句子表示学习方法
- 预训练
  - 了解预训练的基本内容和价值

谢谢各位!