Transformer Plus：未来AI大模型架构设想

刘群 LIU Qun

Huawei Noah's Ark Lab

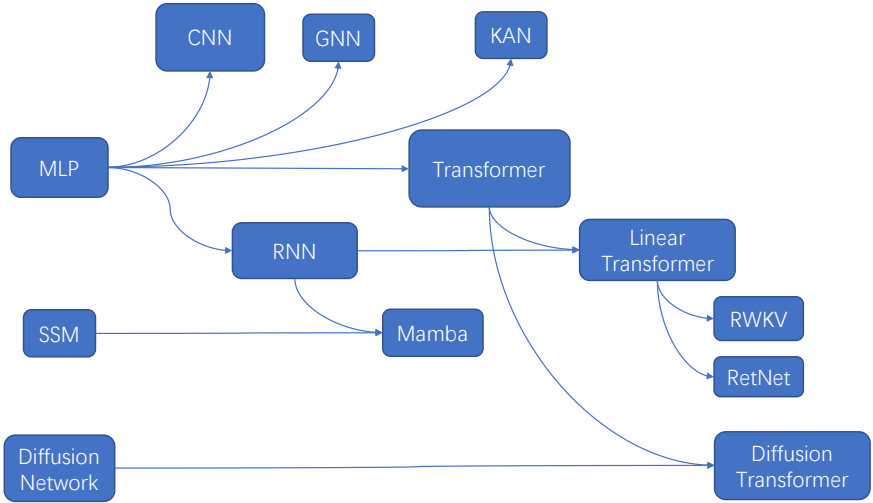
人工智能与算力底座研讨会

2024.05.10-11, 东莞

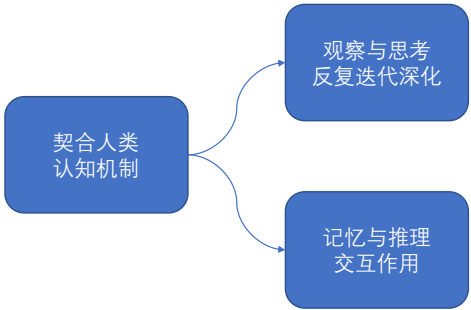
NOAH'S ARK LAB

HUAWEI

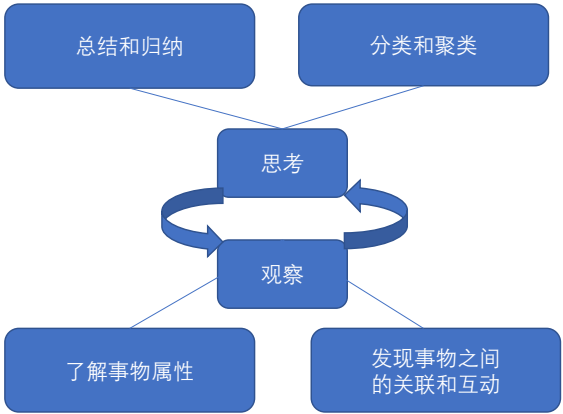
深度学习模型的发展路径



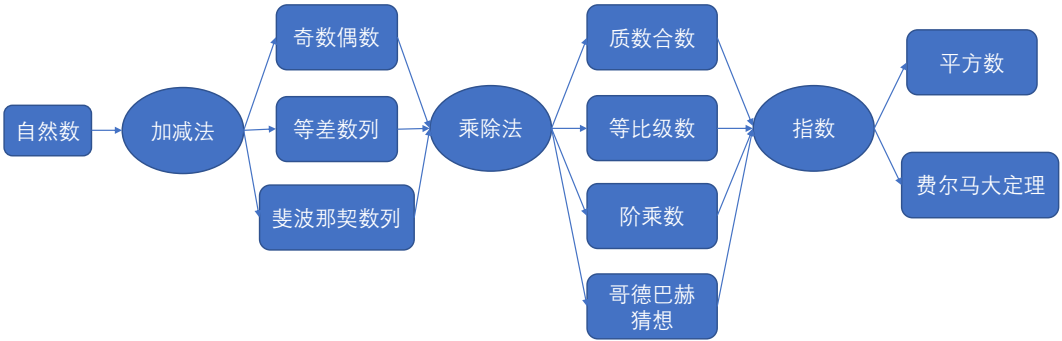
Transformer为什么这么成功



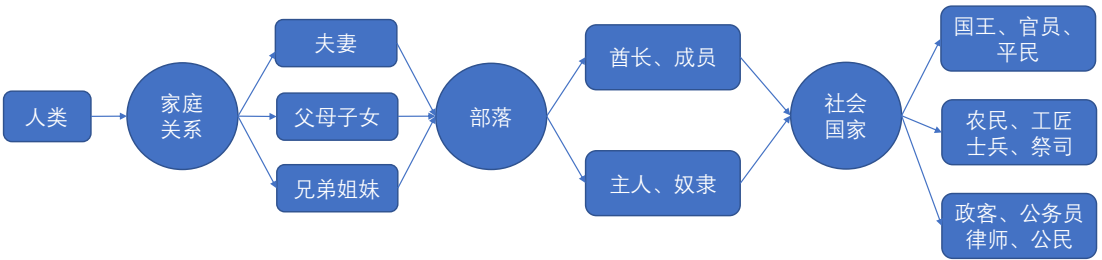
人类认知过程——观察与思考的反复迭代



人类对自然数的认知过程

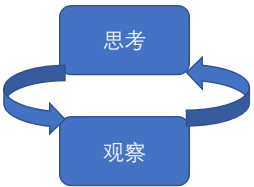


人类对自然语言的认知过程

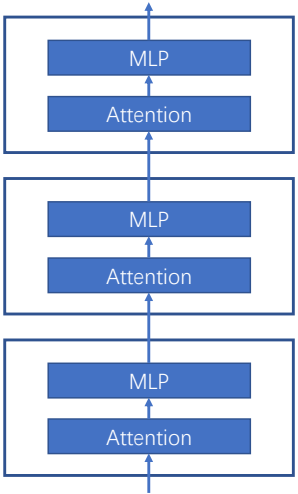


Transformer模拟了观察与思考的迭代过程

MLP层：分类聚类 ➔ 思考



Attention层：建立关联 ➔ 观察



Transformer模拟了记忆与推理的交互过程

- Transformer为每个token保留KV值
  - 记忆可以随着context的增长而无限加长，避免了记忆的弱化问题
- 多层深度网络的迭代，可以很好地学到序列中潜在的复杂结构
  - 很多研究工作表明Transformer可以很好地学到语言的结构信息
  - 结构信息有利于在推理的时候聚焦在相关部分，忽略不相关部分
- Softmax具有极好的选择性，可以准确找到记忆中的相关部分
  - Colins et al., In-Context Learning with Transformers: Softmax Attention Adapts to Function Lipschitzness, arXiv:2402.11639v1

Transformer模型的弱点

- 复杂度：
  - Attention的时空复杂度是context长度的平方
  - MLP的时空复杂度是隐藏层表示维度的平方
- 稀疏性
  - Attention是高度稀疏的，每个token推理只需attend到少数历史token
  - MLP中间层神经元的激活也是高度稀疏的，每次只有少量神经元被激活
  - 不同层之间相似度很高
- 跟符号表示的GAP
  - 人的思考有一部分是用符号进行的
    - 比如一个人熟悉的人，在大脑中一定有神经元直接对应
    - 人在理解句子的时候，大脑中通常会形成一副图像
    - 人脑中有一个海马体，可以保留人的记忆信息，并对时间有感知
  - 现在的深度神经网络（包括Transformer）在进行处理符号的时候还很笨拙，体现在对符号的表示推理都非常不经济，效率很低

未来AI大模型架构设想：Transformer Plus

- Transformer的主体架构不变
- 利用模型的稀疏性，对模型结构进行优化，提高效率，降低复杂度
  - 量化、剪枝
  - 跨层推理（Skip-layer、Early Exit、Mixture-of-Depth）
  - 以存代算
    - Attention层和MLP层都可以kNN相似度检索，找到激活度高的token或者神经元，而忽略其他token和神经元
    - kNN检索复杂度远低于序列长度或表示维度
- 与符号计算的结合
  - 外部结合方式：引入外部记忆单元和符号计算单元
  - 内部结合方式：在Transformer内部引入符号表示单元

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

HUAWEI