中国科学院大学
University of Chinese Academy of Sciences

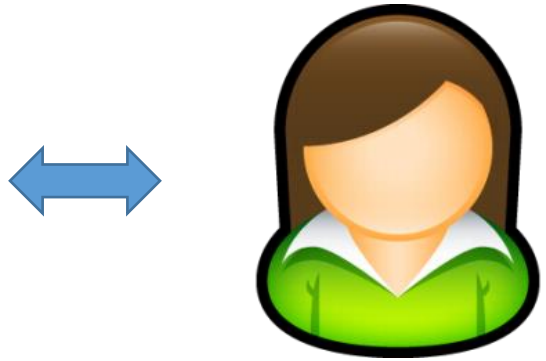
深度学习应用（自然语言处理）

AI DISCOVERY

什么是自然语言处理?

AI DISCOVERY

- 自然语言处理（NLP）是指机器理解并解释人类写作、说话方式的能力
- 自然语言处理的目标是让计算机理解人类的语言，从而弥补人类交流（自然语言）和计算机理解（机器语言）之间的差距



AI DISCOVERY

自然语言处理流程

AI DISCOVERY

◆ 自然语言处理的两大基本步骤是**自然语言理解**和**自然语言生成**

自然语言理解

理解给定文本的含义，将人类的自然语言输入进行分析、计算，产生计算机可以理解和处理的统一形式，需要解决以下几个歧义性

- 词法歧义性：单词有多重含义
- 句法歧义性：句子有多重解析树
- 语义歧义性：句子有多重含义
- 回指歧义性：前后相同的词不同义

自然语言生成

将计算机产生的语义表达结果转化为人类可以读懂的自然语言的过程，大致可以分为三个阶段

- 文本规划：完成结构化数据中基础内容的规划
- 语句规划：从结构化数据中组合语句，来表达信息流
- 实现：产生语法通顺的语句

AI DISCOVERY

自然语言处理技术概览

AI DISCOVERY

NLP基础技术

- 词汇表示和语义分析
- 短语表示和语义分析
- 句子表示和语义分析
- 篇章表示和语义分析

NLP核心技术

- 情感分析
- 信息检索
- 机器翻译
- 机器阅读
- 问答系统
- 语言生成
- 推荐系统
- 聊天和对话

NLP+

- 智能分析
- 搜索引擎
- 商业智能
- 舆情分析

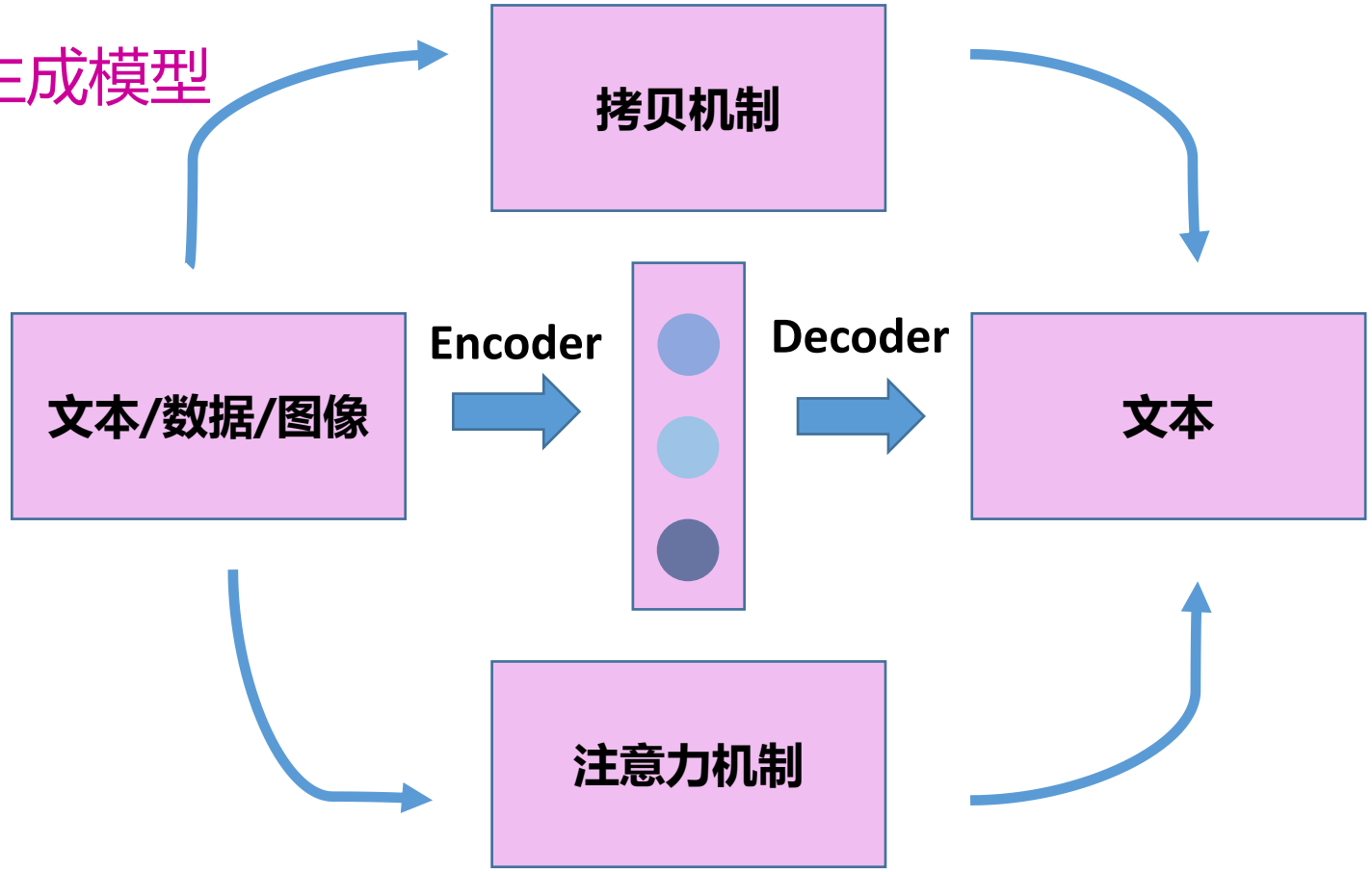
- 大数据
- 云计算
- 机器学习
- 深度学习
- 用户画像
- 领域知识

AI DISCOVERY

文本生成模型

◆ 基于神经网络的生成模型

- 编码器-解码器
- 注意力机制
- 拷贝机制

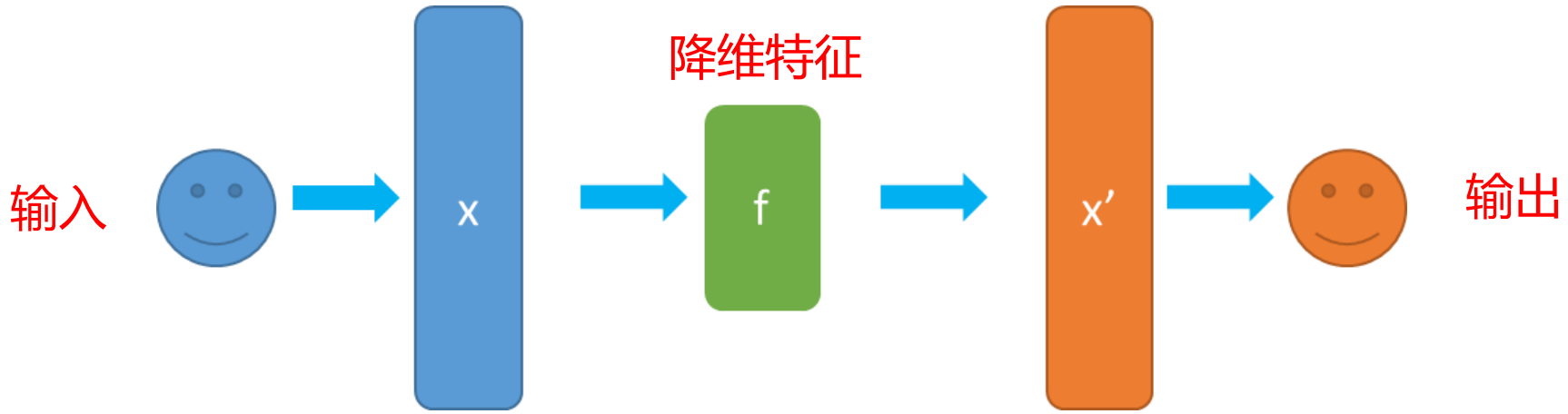


自编码器是什么

AI DISCOVERY

自编码器(AutoEncoder)是一种前馈神经网络，目标是尽可能的让输出与输入一致。

AutoEncoder 使用反向传播进行训练，是无监督模型，主要是用于数据的降维或者特征的抽取上 。



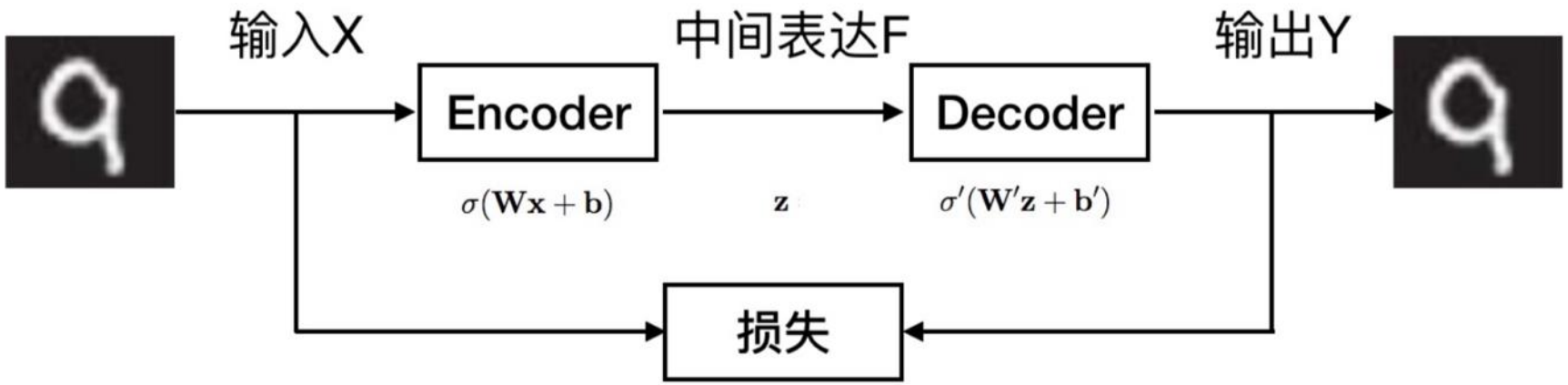
前融合是指在不同模态数据之间进行联合特征抽取，而自编码模型主要用于数据降维或者特征抽取，正是因为自编码模型具有提取特征的能力，所以可用于前融合

AI DISCOVERY

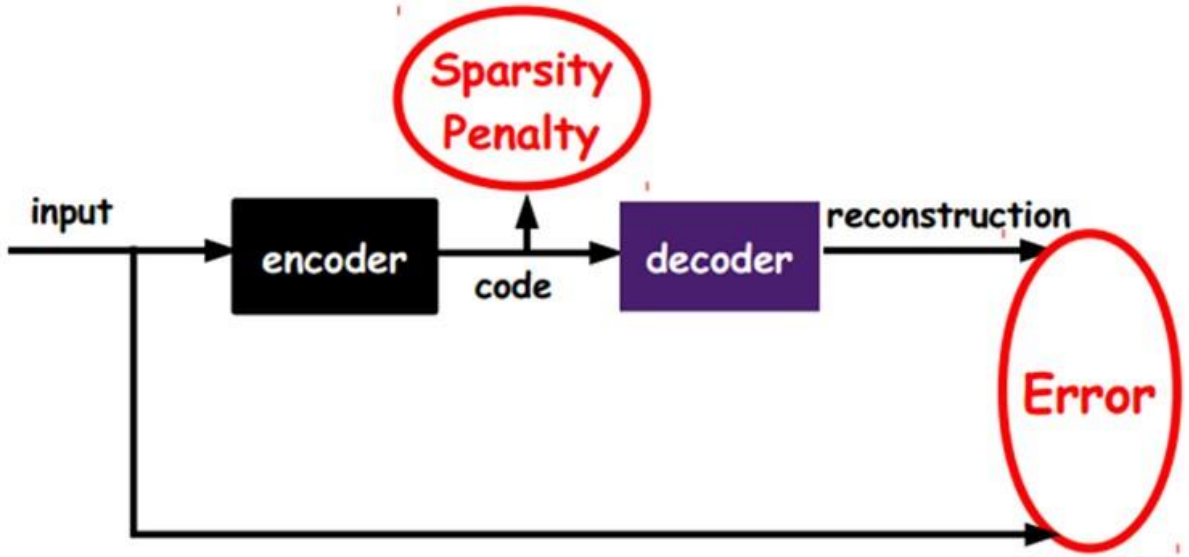
自编码器原理

AI DISCOVERY

- AutoEncoder 通常包含两个部分:编码器和解码器。
  - 编码器和解码器可以是单层的也可以是多层的，多层的编码器和解码器往往性能更好。
  - 输入X经过编码器得到中间表示层F，又称为编码(encode)，然后 F 经过解码器得到输出 Y，两个过程分别称为编码过程和解码过程。
  - AutoEncoder 的目标函数可以表示为输出和输入的差值最小。



什么是稀疏自编码



- input: X code: $h = W^T X$

- loss: $L(X;W) = \|Wh - X\|^2 + \lambda \sum_j |h_j|$ ← L1正则

✓ 稀疏自编码器（Sparse AutoEncoder）可以约束中间表达尽可能的稀疏，能够学习到更加有用的特征。

✓ 如果在AutoEncoder的基础上加上L1正则限制（L1主要是约束每一层中的节点中大部分都要为0，只有少数不为0，这就是Sparse名字的来源），可以得到Sparse AutoEncoder。