

AutoEncoder 与 Bow 学习笔记

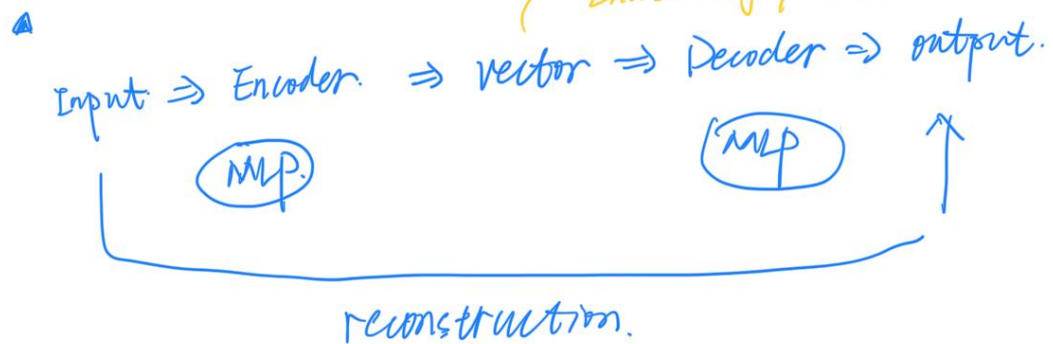
AutoEncoder:

Auto-Encoder.

- 是自监督学习的一种，早于大模型方法提出。

Encoder 和 输出过程:

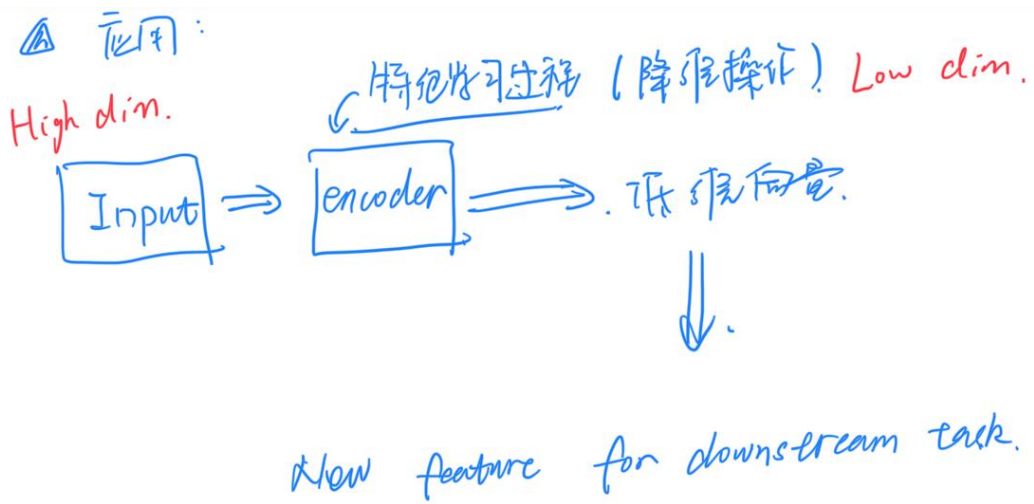
(Embedding / Representation.



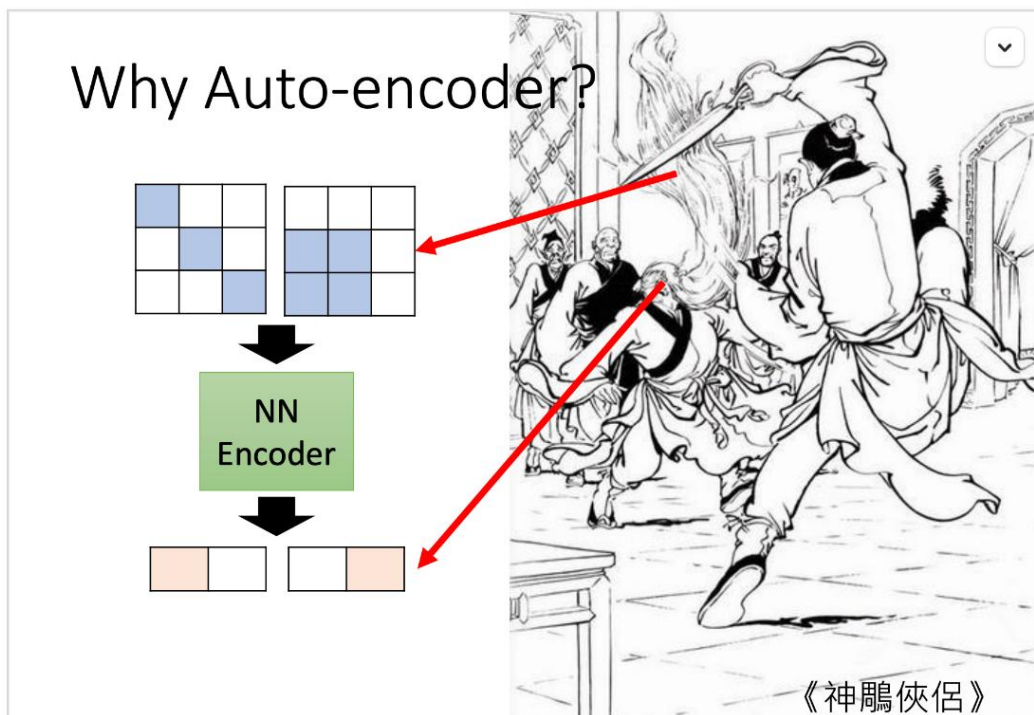
目标: Input 经过 Encoder, Decoder To

生成 Output 要求与 Input 越接近越好.

最小化输入和输出差异.



Encoder 的核心: 化繁为简的过程.
提取关键特征.



自编码器在高维数据降维和特征学习上表现出强

大能力。

主要应用：

主要应用

- 降维替代 PCA（特别是非线性场景）。
- 特征学习：从无标签数据中学习表示。
- 数据压缩：图像/语音数据的低维表示。
- 生成模型（扩展后形成 VAE、GAN 等）。
- 异常检测：输入与重建差异大，说明该样本可能是异常。

参考资料：

https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/auto_v8.pdf

参考视频：<https://www.youtube.com/watch?v=3oHlf8-J3Nc>

BoW:

核心思路:

1. 将-FS子串看作 Bag.
2. 只关心词频, 不考虑词序和语法结构.
(TF / TF-IDF).

应用场景:

1. 文本分类. (如: 垃圾邮件识别、情感分析).
2. 信息检索.

AE 与 BoW 结合:

BoW 是词频统计器, 不考虑文本词语与语法结构,

AE 是一种神经网络结构, 可以学习到隐含的语义特征的向量表示, 可以保留更多的上下文语义信息。

两者的信息利用方式不同, BoW 简单高效, 适合传统的机器学习方法。AutoEncoder 适合深度学习方法, 用于文本降维、去噪、特征提取。

结合方式之一:

将 BoW 的显式特征 (词频) 和 AutoEncoder 的隐式特征 (语义嵌入) 拼接在一起, 得到更丰富的表示。