国科大. 深度学习：期末复习知识点总结笔记_国科大 张新峰 深度学习 - CSDN 博客

- 卷积核在图片上如何计算，padding，特征图大小，参数量
- 图卷积网络，给定几个点的状态，权重矩阵，进行一次图卷积，求结果
- 几种激活函数的对比
- sigmoid：
    - Sigmoid 函数的输出映射在 (0,1) 之间，单调连续，输出范围有限，优化稳定，可以用作输出层。它在物理意义上最为接近生物神经元。
    - 容易产生梯度消失，导致训练出现问题。
- tanh：
    - 比 Sigmoid 函数收敛速度更快。
    - 相比 Sigmoid 函数，其输出以 0 为中心。
    - 还是没有改变 Sigmoid 函数的最大问题——由于饱和性产生的梯度消失。
- ReLU：
    - ReLU 在 SGD 中能够快速收敛。据称，这是因为它线性、非饱和的形式。
    - 可以更加简单的实现。
    - 有效缓解了梯度消失的问题 。
    - 随着训练的进行，可能会出现神经元死亡，权重无法更新的情况。
- 全连接网络，如何求损失函数相对于网络参数的梯度
- 梯度消失爆炸的原因与解决
    - 在反向传播过程中需要对激活函数进行求导，如果导数大于 1，那么随着网络层数的增加梯度更新将会朝着指数爆炸的方式增加

这就是梯度爆炸。同样如果导数小于 1，那么随着网络层数的增加梯度更新信息会朝着指数衰减的方式减少这就是梯度消失。因此，梯度消失、爆炸，其根本原因在于反向传播训练法则，属于先天不足。

- 另一方面，如果选择 sigmoid 激活函数，导数小于 1，也容易导致梯度消失
- 解决方案：
    - 改激活函数：ReLU，leak ReLU
    - Batchnorm，批规范化，通过规范化操作将输出信号 x 规范化到均值为 0，方差为 1 保证网络的稳定性。加速网络收敛速度，提升训练稳定性的效果。

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$$

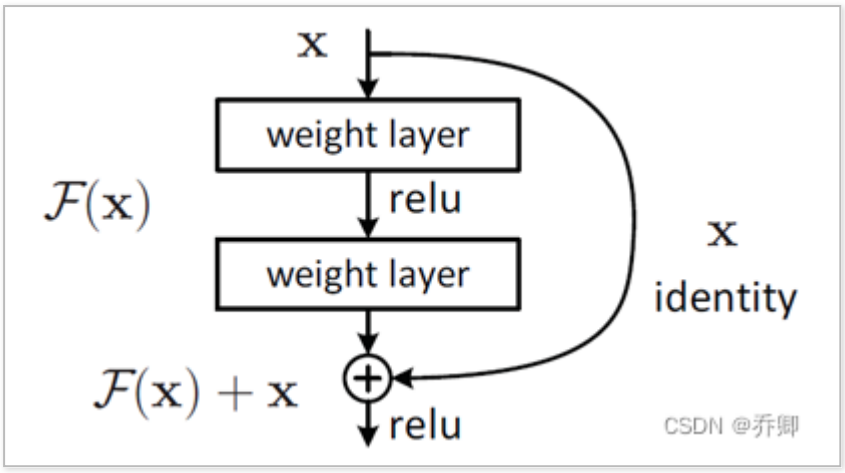
$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$$

CSDN @乔卿

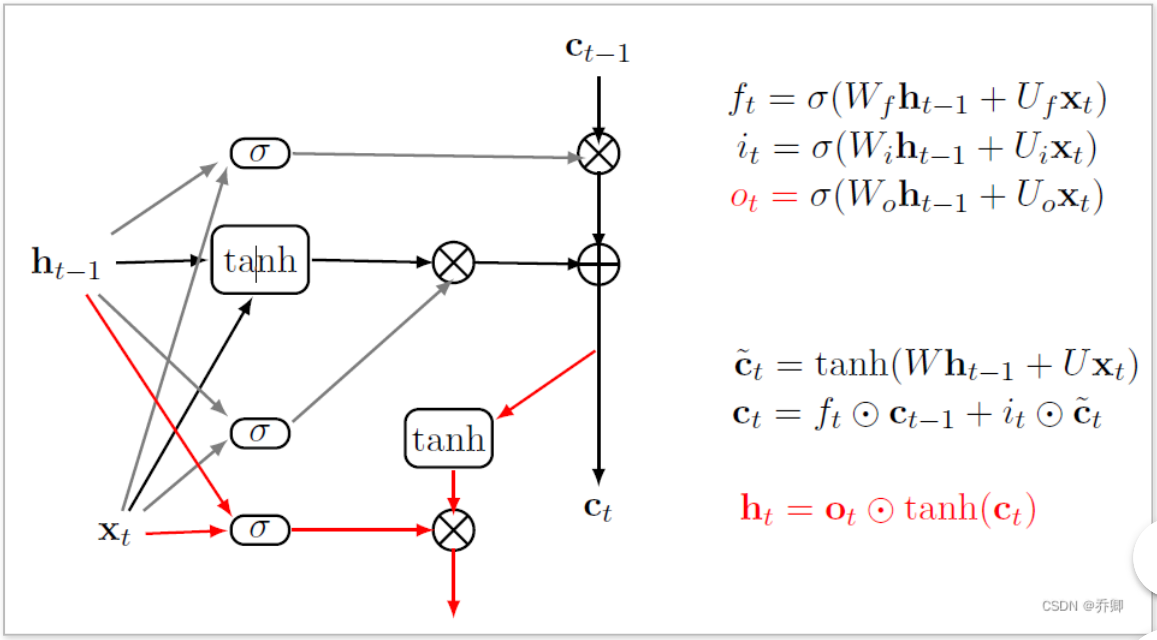
- - - 残差结构：网络深度增加时，网络准确度出现饱和，甚至出现下降。短路连接。



$$\frac{\partial loss}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \frac{\partial x_L}{\partial x_l} = \frac{\partial loss}{\partial x_L} \cdot \left(1 + \frac{\partial}{\partial x_L} \sum_{i=l}^{L-1} F(x_i, W_i)\right)$$

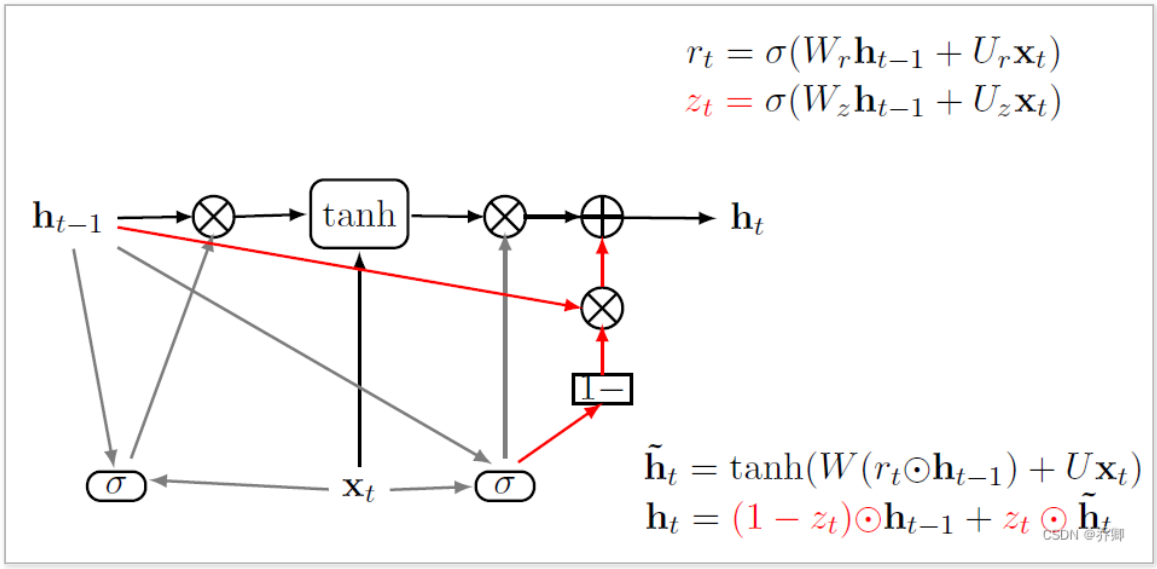
CSDN @乔卿

- GRU、LSTM，解决什么问题
  - 解决的问题：RNN 在许多阶段传播的梯度往往会消失（大部分时间）或爆炸（相对很少）。与短期交互作用相比，RNN 难以建模长期依赖关系。另外，RNN 并不总是容易训练。LSTM 可以解决梯度消失问题！
  - LSTM：输入门、输出门、遗忘门

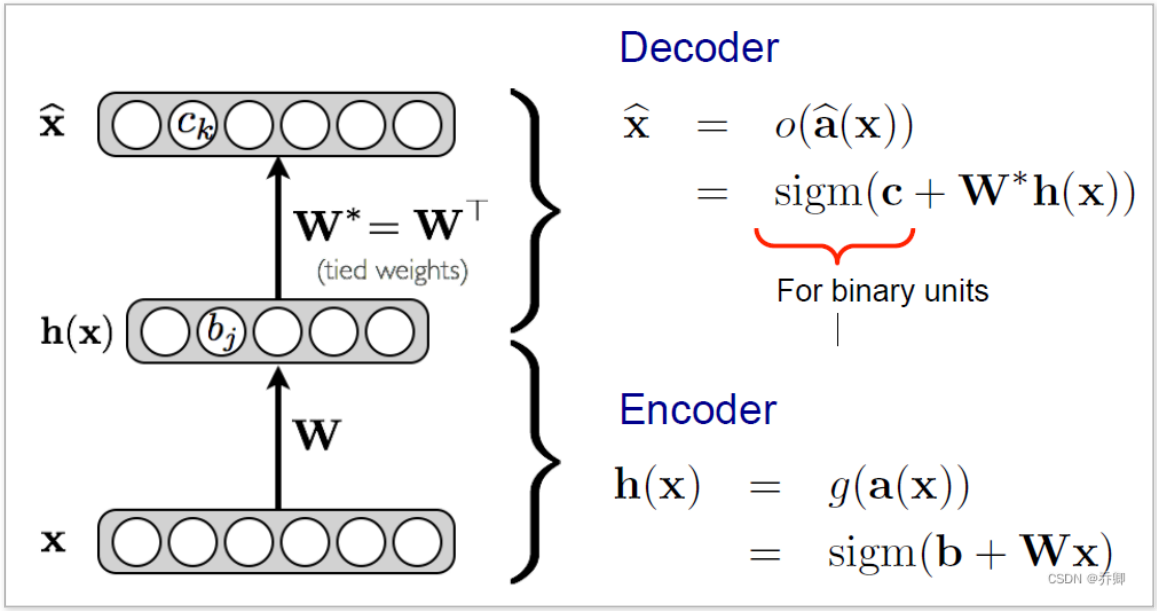


牢牢记住：Wh+Ux

- - GRU：复位门、更新门



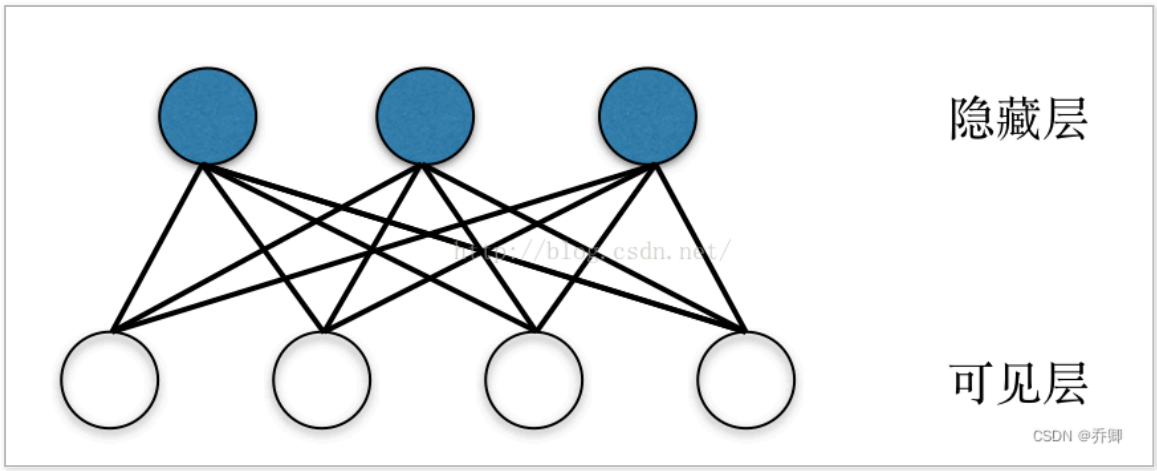
- RBM、DBN 与 GAN、VAE 对比，优劣势与特点
  - AE：自编码机



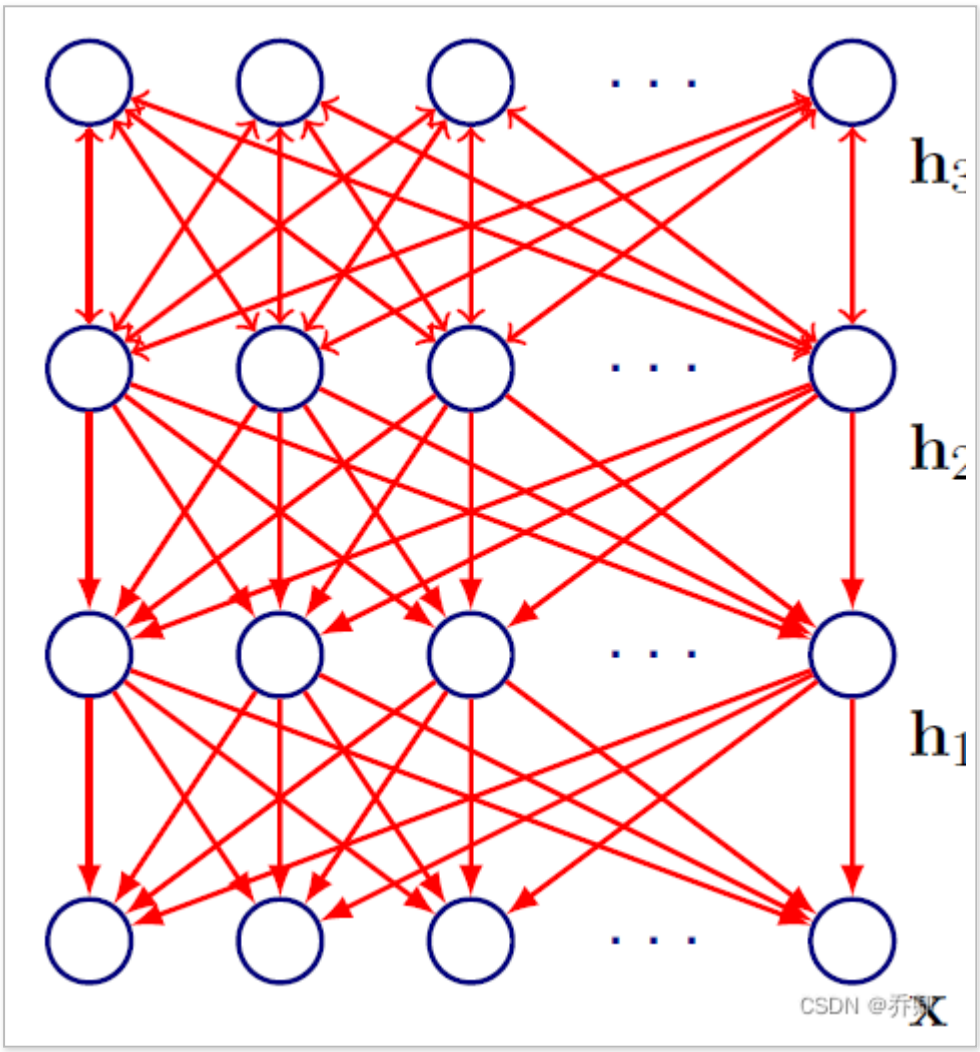
过程中降低了特征维度！学习低维表示，希望能无损地重构，重构误差最小

- - RBM 受限玻尔兹曼机、DBN 深度置信网络

RBM 可以学习数据内部特征，拟合离散分布，基于能量模型



DBN 逐层无监督训练 RBM，最后有监督微调



- GAN

$$\min_G \max_D V(D,G)$$

$$V(D,G) = \mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{z\sim q(z)}\left[\log\left(1 - D(G(z))\right)\right]$$

- VAE

VAE 模型是一种包含隐变量的生成模型，它利用神经网络训练得到两个函数（也称为推断网络和生成网络），进而生成输入数据中不包含的数据。基于概率。

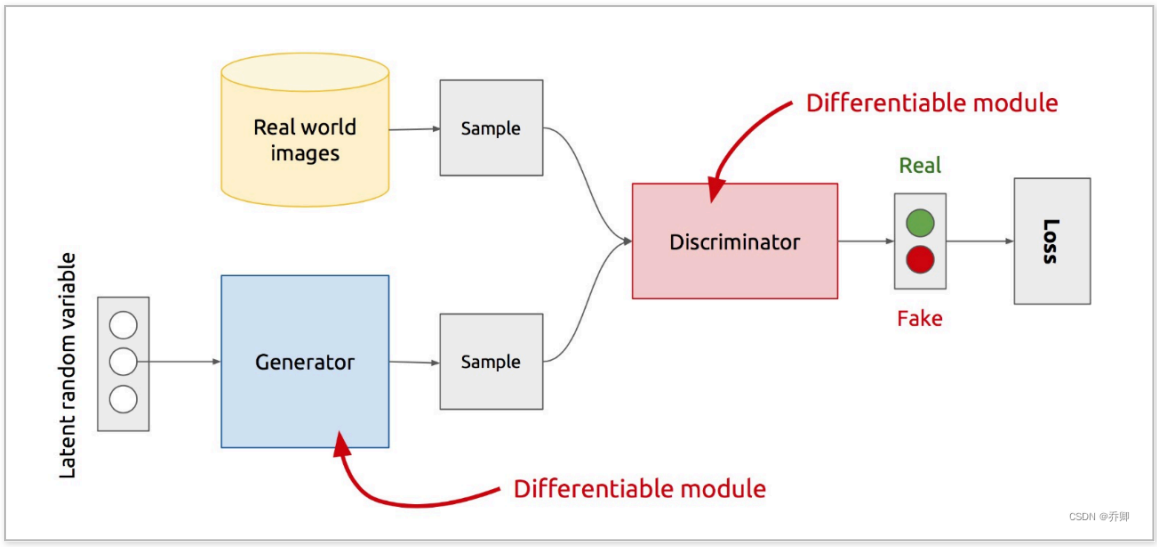
VAE 中隐藏层服从高斯分布，AE 中的隐藏层无分布要求。

训练时，AE 训练得到 Encoder 和 Decoder 模型，而 VAE 除了得到这两个模型，还获得了隐藏层的分布模型（即高斯分布的均值与方差）

AE 只能重构输入数据 X，而 VAE 可以生成含有输入数据某些特征与参数的新数据。

相比于传统机器算法，GAN 有三方面的优势：

- 首先 GAN 模型的表现效果更好，生成清晰的样本；
- 第二 GAN 框架可以训练任何一种生成器网络；
- 第三 GAN 适用于一个变量的随机发生概率不可计算的情况。



- 强化学习基本思想、基本要素、应用场景

基本思想：智能体，环境，状态，动作，奖励，最大化期望的奖励，监督学习与强化学习相结合

马尔可夫决策过程的定义：$(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbb{P}, \gamma)$ CSDN @乔卿 ，状态，动作，奖励，转移概率，奖励衰减因子

状态估值函数的贝尔曼最优，贝尔曼方程：

$$Q_{i+1}(s,a) = \mathbb{E}\left[r + \gamma \max_{a'} Q_i(s',a') | s,a\right]$$

CSDN @乔卿

，别忘了权重衰减因子！

- 几种注意力

注意力机制就是对输入权重分配的关注，最开始使用到注意力机制是在编码器 - 解码器 (encoder-decoder) 中， 注意力机制通过对编码器所有时间步的隐藏状态做加权平均来得到下一层的输入变量。

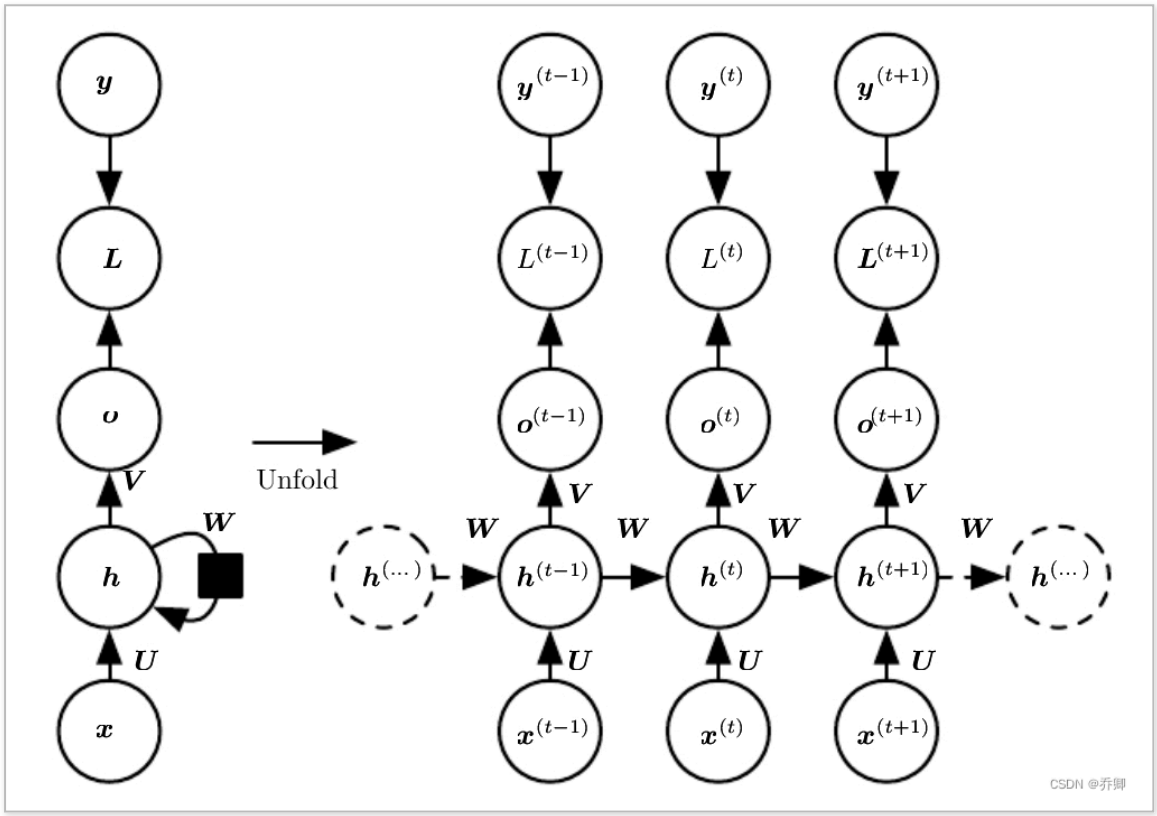
- - soft attention：易于实现：在输入位置上生成分布，重新加权特性并作为输入馈送，使用空间变压器网络关注任意输入位置

- hard attention：依概率选择一个，关注单个输入位置，无法使用梯度下降！需要强化学习！

- 循环神经网络：RNN 的结构、优化

$$s^{(t)} = f(s^{(t-1)}, x^{(t)}; \theta)$$

注意三个权重矩阵！U、V、W



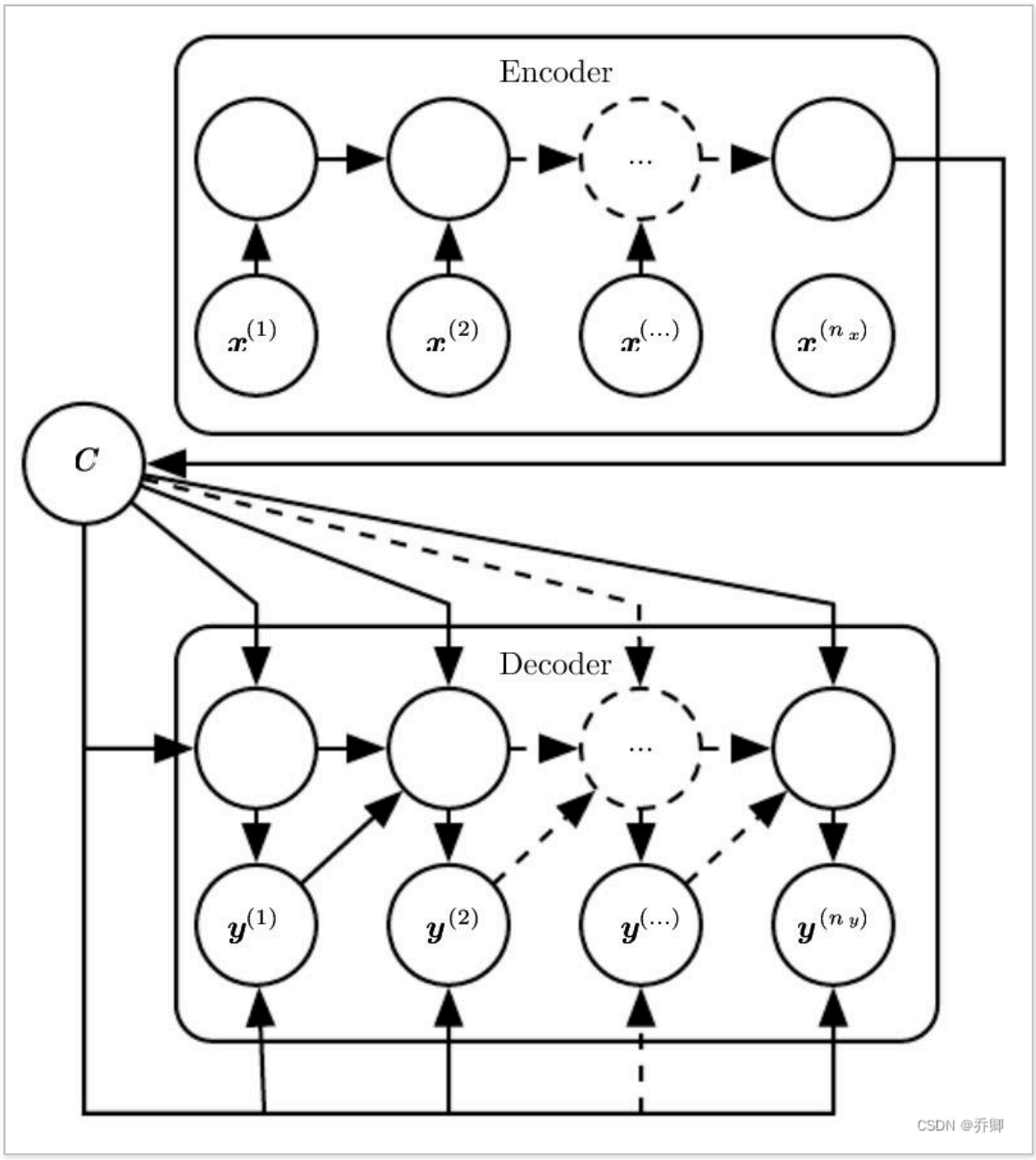
同时取决于输入与前一时刻的输出

- 反向传播：BPTT

损失与梯度都是对所有的 t 相加！

U、V、W 是共享的！

- 用于机器翻译：



全文完