



哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

深度学习体系结构（实验）

实验三 基于LSTM的MNIST手写数字识别加速

实验目的

- 加深对循环神经网络RNN基本原理的理解
- 了解LSTM识别手写数字的基本原理
- 掌握使用HLS设计实现LSTM硬件加速器的基本流程和方法

实验内容

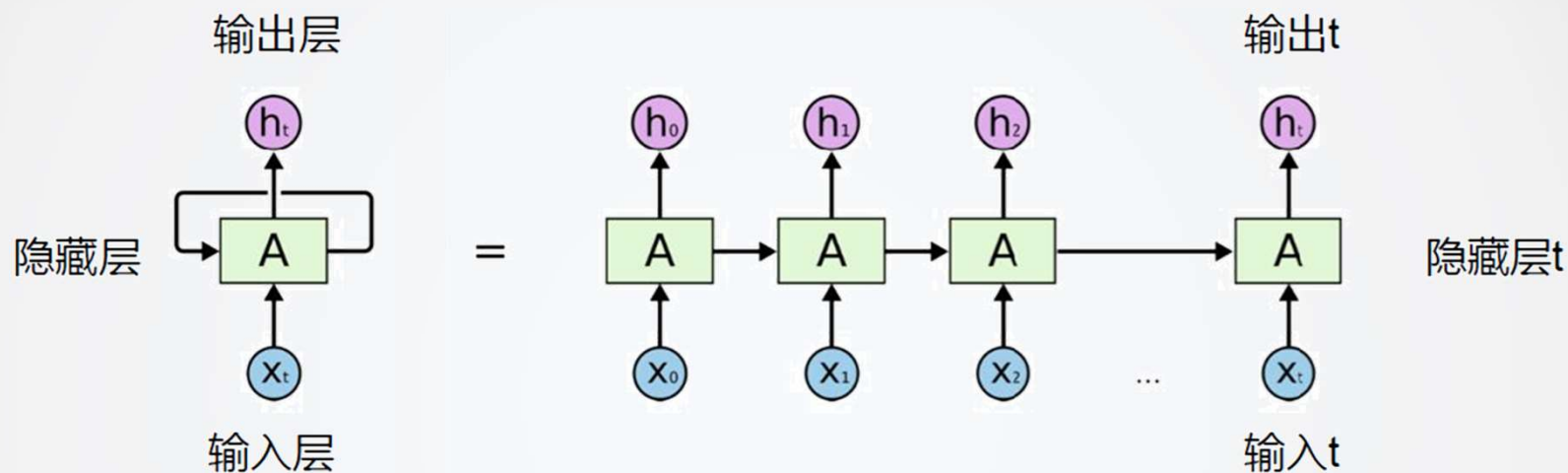
- 使用HLS编写RNN，包括LSTM Cell和向量的Sigmoid、Tanh、加法等
- 构建Block Design电路图，生成Overlay并上板测试
- 观察并对比分析软硬件推导的差别

实验原理 — 循环神经网络简介

RNN (Recurrent Neural Network)

- 使用带有**自反馈**的神经元
- 可处理任意长度的输入序列，以及相互依赖、可变长输入的任务
 - 视频（由连续的图片组成，长度不定）、文章（词义依赖于上下文，长度不定）
 - 大小不同的图片、时长不一的视频、长短各异的句子，etc
- 已广泛应用在语音处理、图像处理、自然语言处理等领域

实验原理 — RNN网络结构

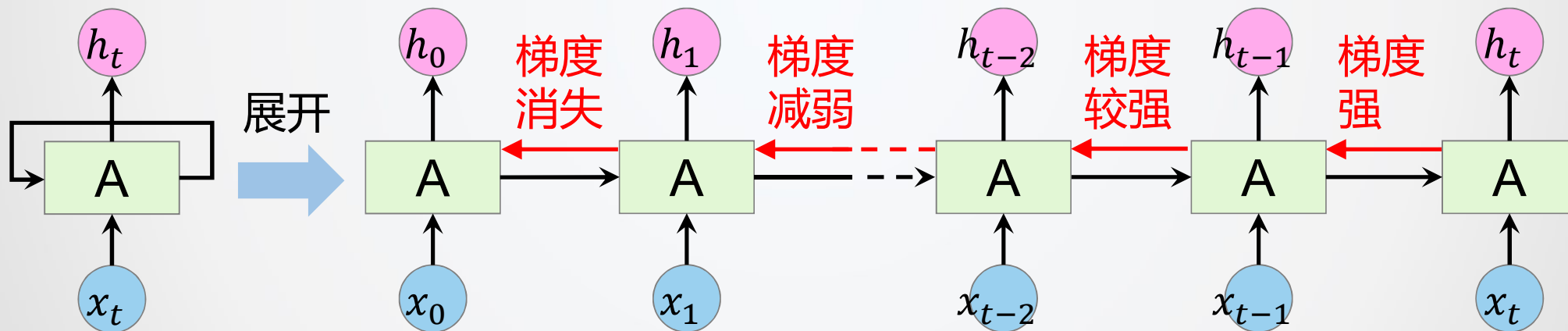


- 给定输入序列 $\{x_1, x_2, \dots, x_t, \dots, x_n\}$, 每单位时间向RNN输入一个 x_i
- 输出计算公式:
$$h_t = \begin{cases} 0 & t = 0, \\ f(h_{t-1}, x_t) & otherwise. \end{cases}$$

实验原理 — 梯度消失问题

- 训练时，使用通过时间的反向传播（Back Propagation Through Time, BPTT）算法

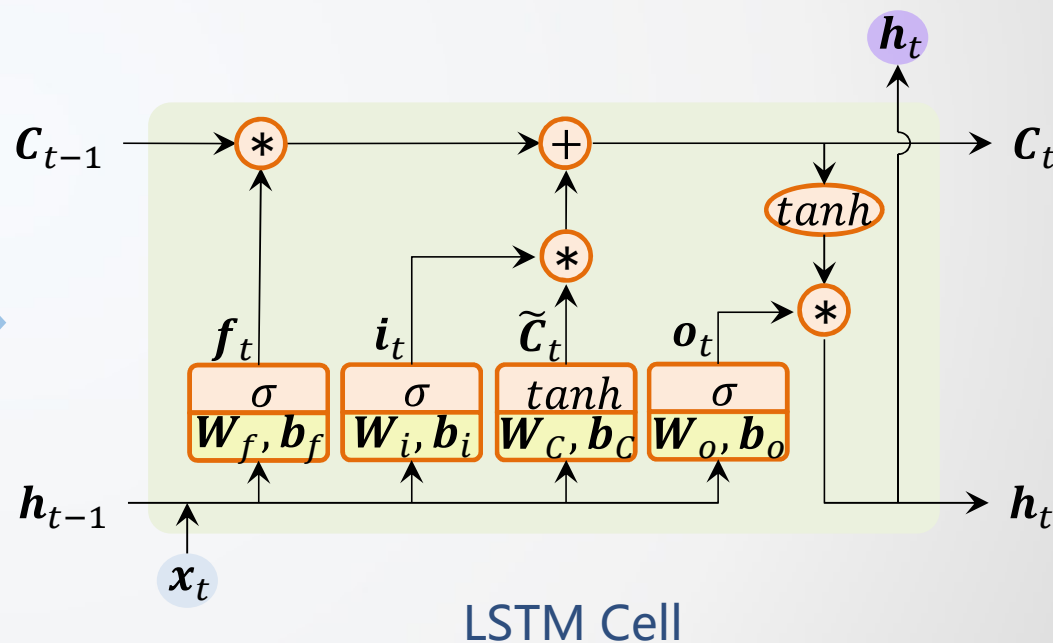
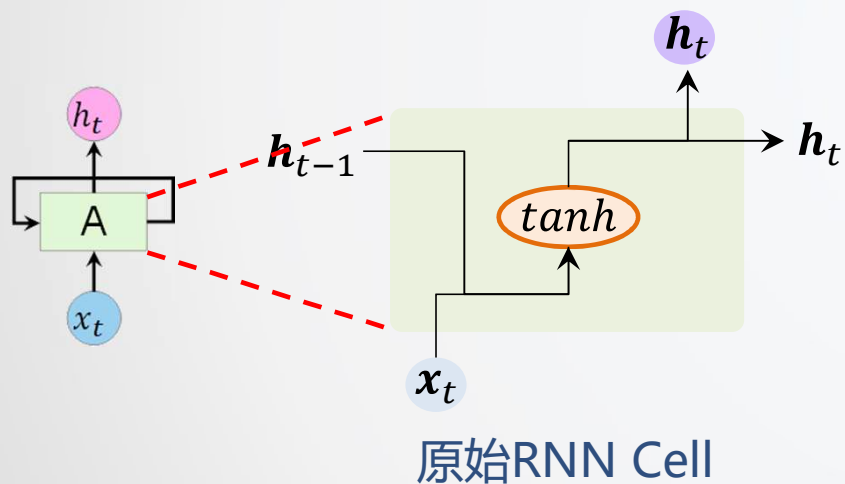
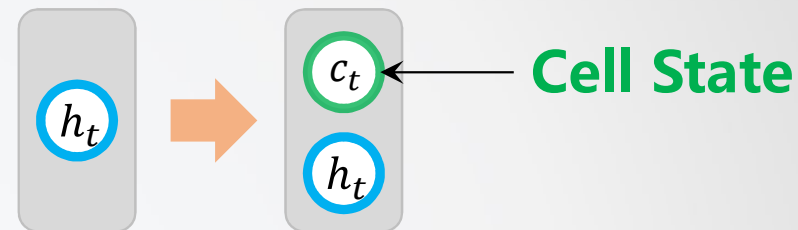
□ 梯度计算使用链式求导：
$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial y_t} \cdot \frac{\partial y_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial y_{t-1}} \cdot \frac{\partial y_{t-1}}{\partial h_{t-1}} \cdots \frac{\partial y_0}{\partial h_0} \cdot \frac{\partial h_0}{\partial W}$$



实验原理 — 长短期记忆

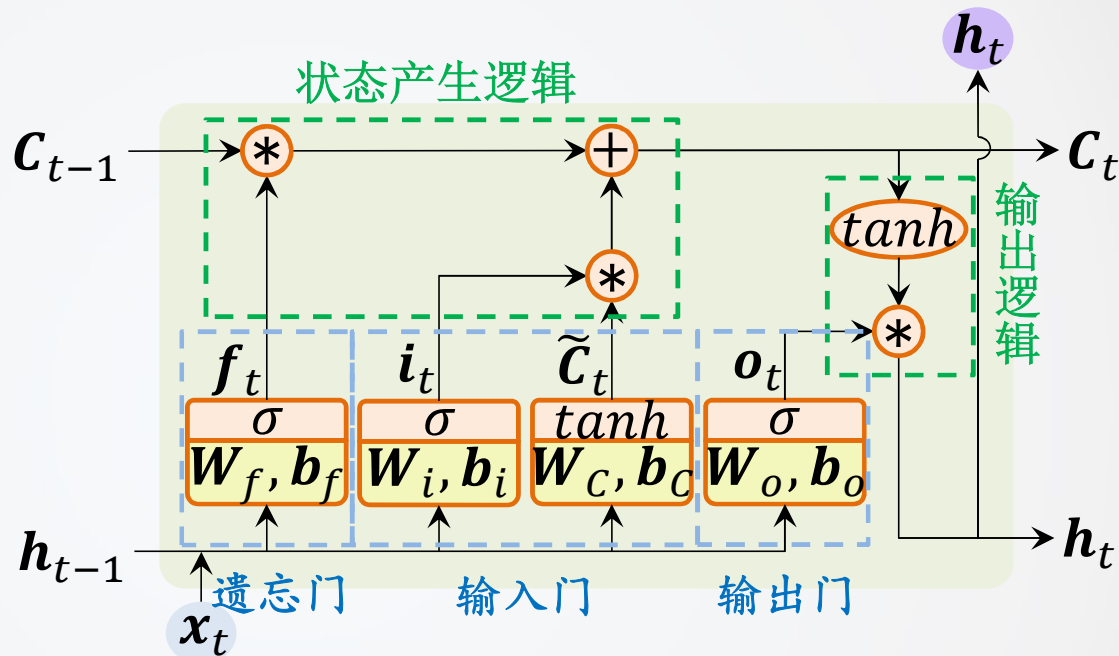
LSTM (Long Short-Term Memory)

- 在原始RNN的基础上，增加**单元状态** (Cell State) 以存储长期信息



实验原理 — LSTM Cell

- 遗忘门
- 输入门
- 输出门
- 状态产生逻辑
- 输出逻辑

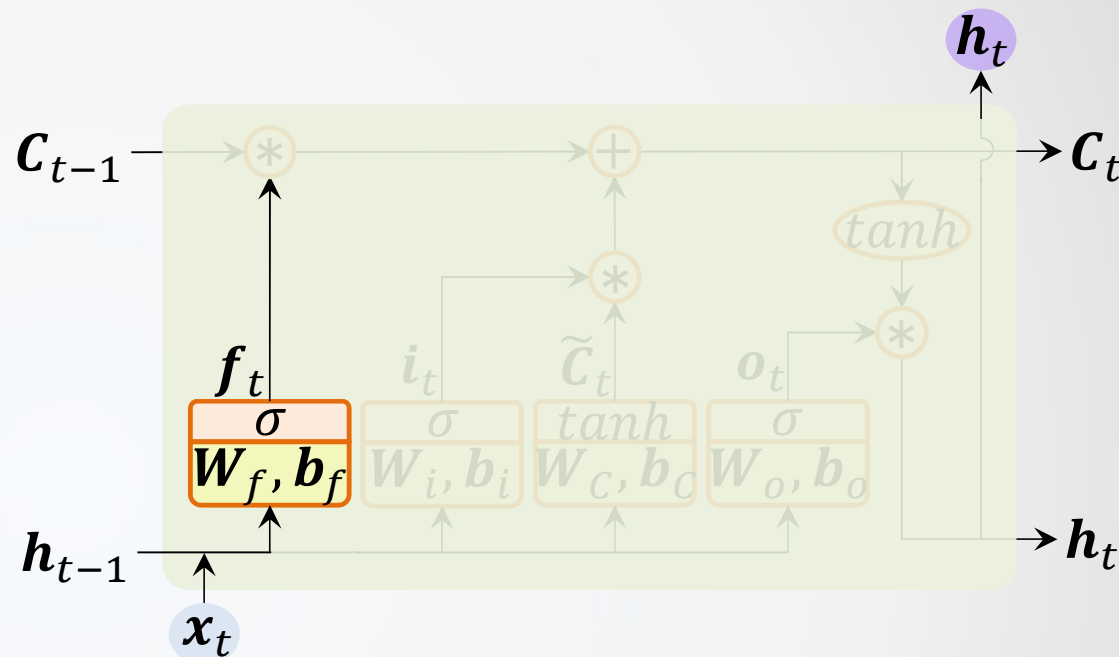


实验原理 — LSTM Cell

遗忘门 (Forget Gate)

- **作用**: 决定记住多少上一时刻的 Cell State (C_{t-1})
- **输入**: 上一时刻的输出 h_{t-1} 、当前时刻的单元输入 x_t
- **输出**: 遗忘门控制向量 f_t
- **计算公式**:

$$f_t = \text{sigmoid}(W_f \cdot [h_{t-1}, x_t] + b_f)$$



实验原理 — LSTM Cell

输入门 (Input Gate)

■ **作用：** 决定要在当前时刻的Cell State (C_t)中存储什么信息

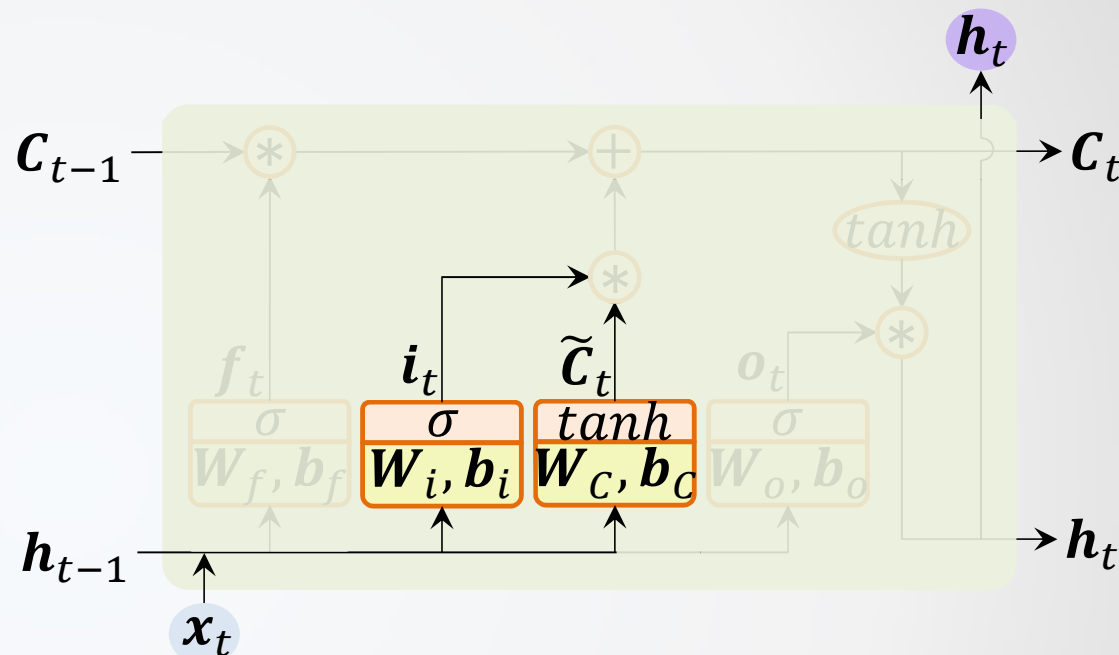
■ **输入：** 上一时刻的输出 h_{t-1} 、当前时刻的单元输入 x_t

■ **输出：**

□ 输入门控制向量 i_t — 决定更新哪些信息

□ 候选状态向量 \tilde{C}_t — 提供待更新的信息

■ **计算公式：**
$$\begin{cases} i_t = \text{sigmoid}(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \end{cases}$$

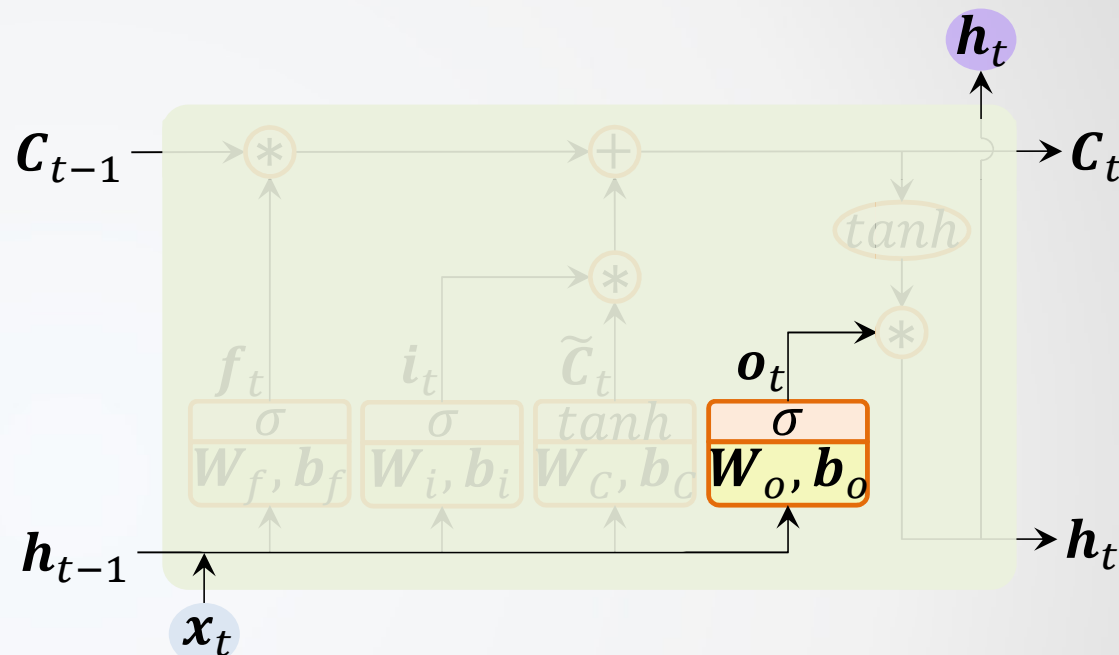


实验原理 — LSTM Cell

输出门 (Output Gate)

- **作用**: 决定t时刻输出哪些信息
- **输入**: 上一时刻的输出 h_{t-1} 、当前时刻的单元输入 x_t
- **输出**: 输出门控制向量 o_t
- **计算公式**:

$$o_t = \text{sigmoid}(W_o \cdot [h_{t-1}, x_t] + b_o)$$



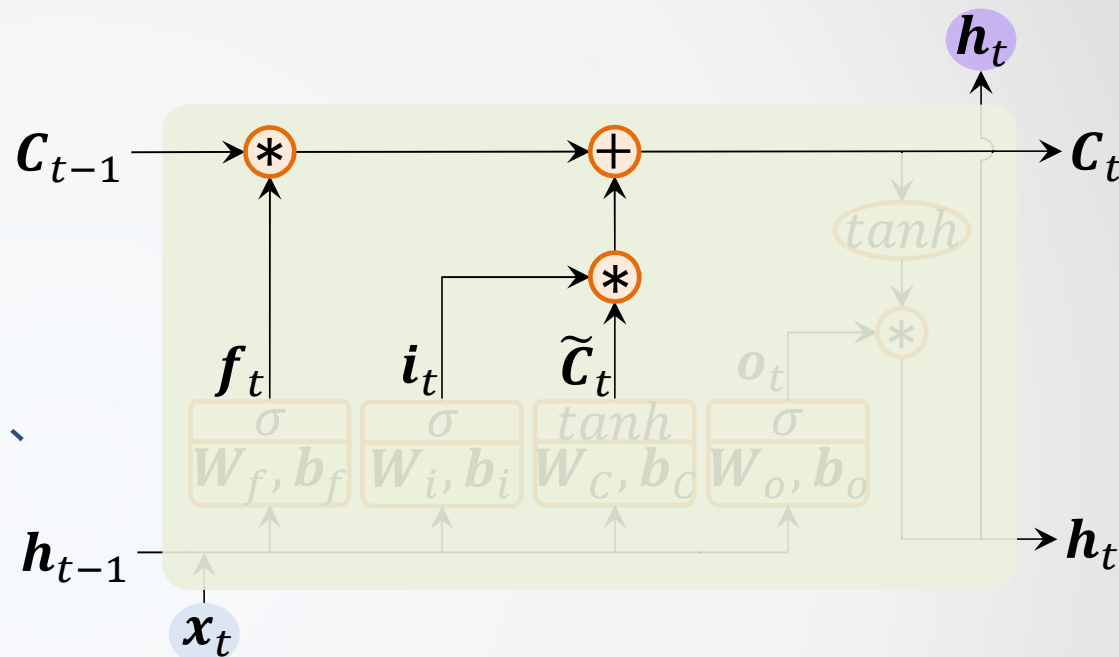
实验原理 — LSTM Cell

状态产生逻辑

- **作用：**产生Cell的新状态 C_t
- **输入：**上一时刻的状态 C_{t-1} 、遗忘门控制向量 f_t 、输入门控制向量 i_t 、候选状态向量 \tilde{C}_t
- **输出：**新状态向量 C_t
- **计算公式：**

$$C_t = C_{t-1} * f_t + i_t * \tilde{C}_t$$

- **Hadamard Product:** 若向量 $\mathbf{a} = \{a_1, a_2, \dots, a_n\}$, $\mathbf{b} = \{b_1, b_2, \dots, b_n\}$, 则 $\mathbf{a} * \mathbf{b} = \{a_1 b_1, a_2 b_2, \dots, a_n b_n\}$

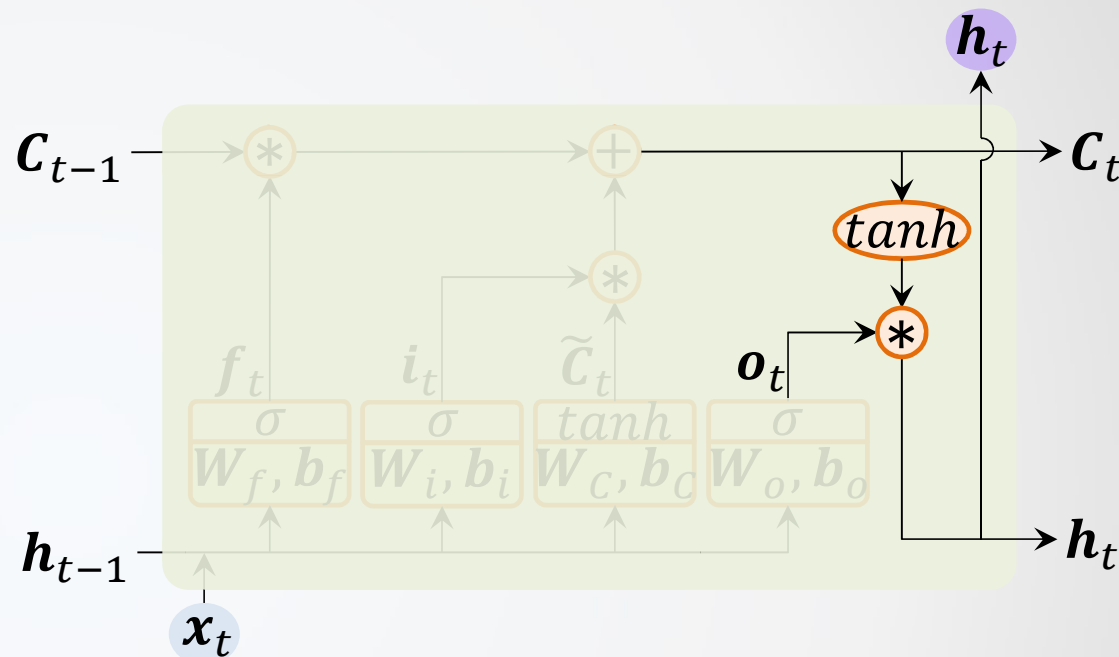


实验原理 — LSTM Cell

输出逻辑

- **作用：**产生Cell的输出 h_t
- **输入：**输出门控制向量 o_t 、当前的状态向量 C_t
- **输出：**当前时刻的输出向量 h_t
- **计算公式：**

$$h_t = o_t * \tanh(C_t)$$

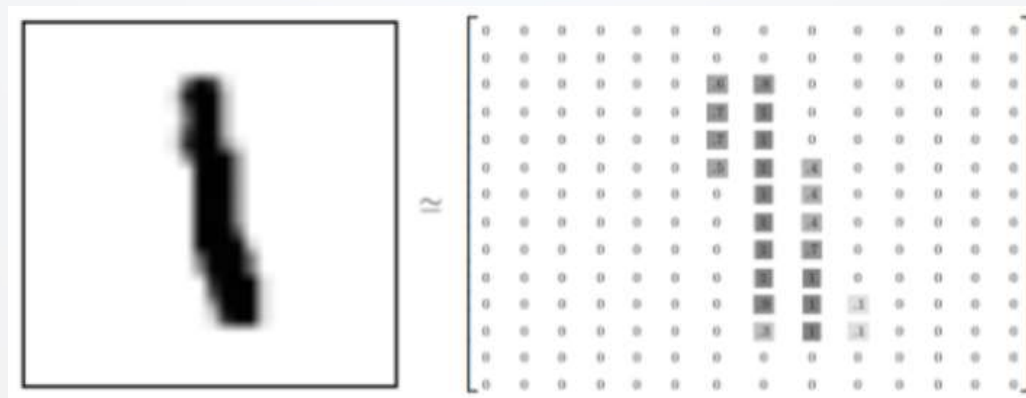


实验原理 — MNIST数据集

<http://yann.lecun.com/exdb/mnist/>

文 件	内 容
train-images-idx3-ubyte.gz	训练集图片（训练集55000张，验证集5000张）
train-labels-idx1-ubyte.gz	训练集图片所对应的数字标签
t10k-images-idx3-ubyte.gz	测试集图片（10000张）
t10k-labels-idx1-ubyte.gz	测试集图片所对应的数字标签

□ 每张灰度图片都是一个 28×28 的二维矩阵



实验原理 — LSTM识别手写数字

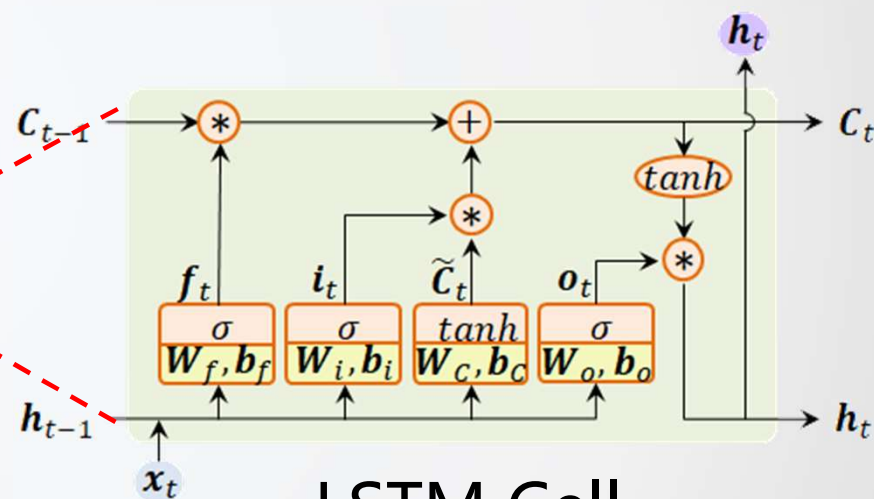
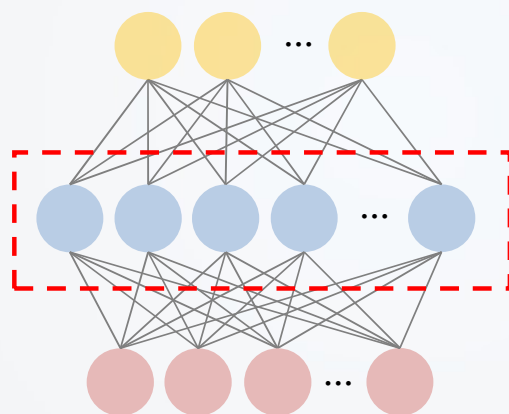
RNN网络结构



输出层 - 10

隐藏层 - 128
(num_units)

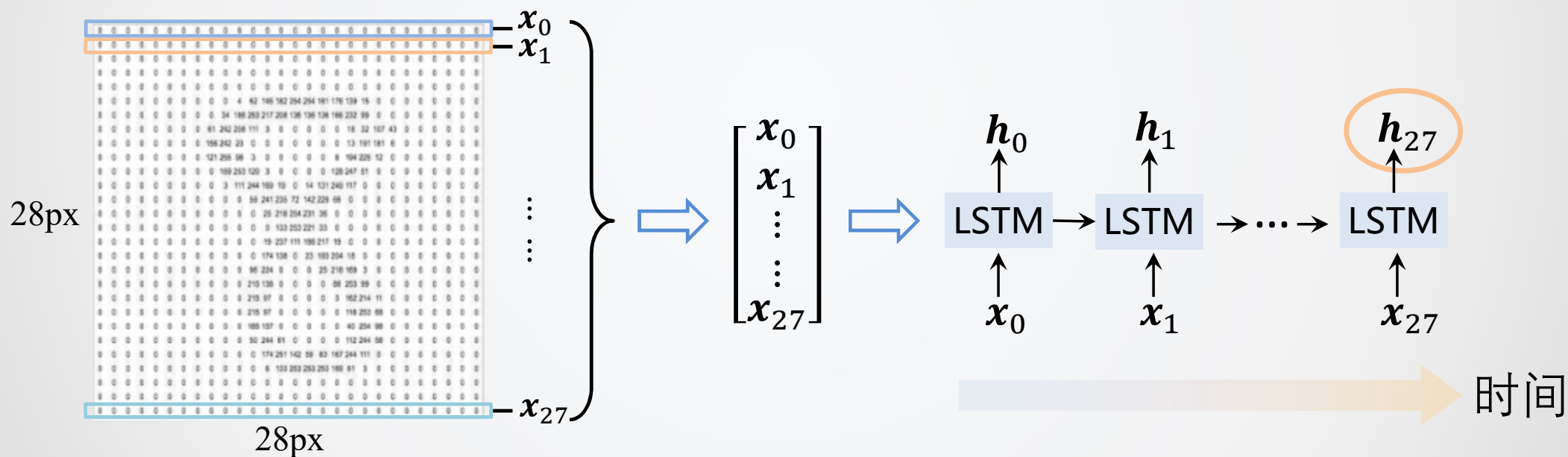
输入层 - 28



LSTM Cell

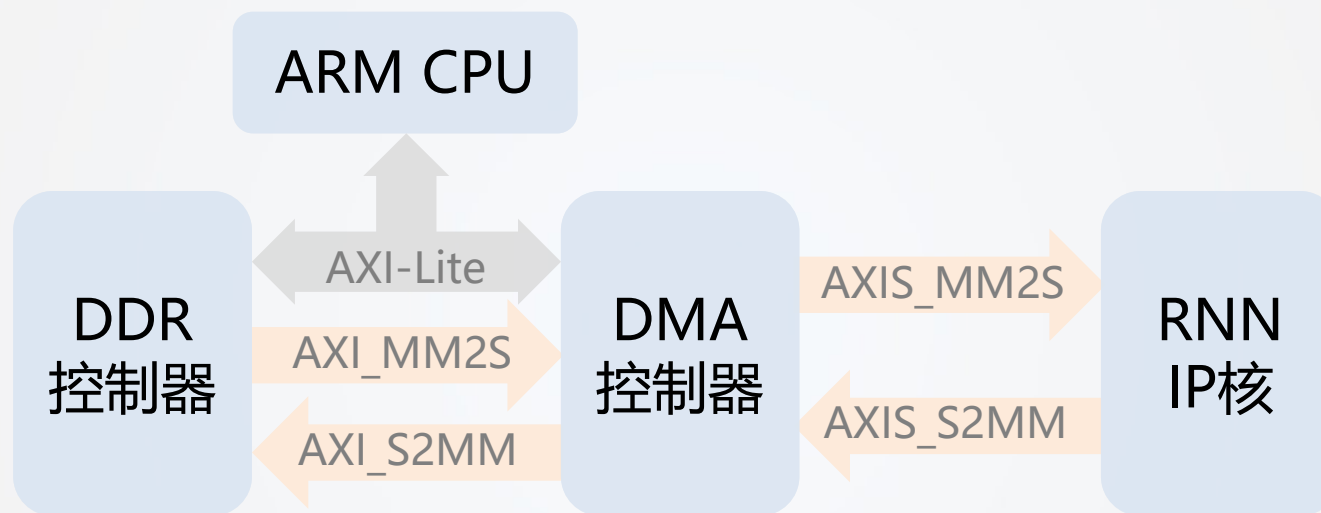
实验原理 — LSTM识别手写数字

- 将图片看作一个包含28个行向量的序列 $\{x_0, x_1, \dots, x_{27}\}$
- 每次向RNN输入一个行向量，共输入28次



实验原理 — RNN加速器架构

■ 采用完全硬件化方法设计



- ARM CPU: 系统初始化、数据预处理
- RNN IP核: 加速RNN前向推导——主要算力
- DMA控制器: 实现PS与PL的快速数据传输
- DDR控制器: 处理数据读写请求

实验步骤

1. 使用HLS编写RNN IP核
2. CSim (需先打开仿真开关)
3. 综合、打包生成IP核 (关闭仿真开关)
4. 构建Block Design、生成并导出Overlay
5. 上板测试

```
// C仿真开关  
#define CSIM_ON 1 // 仿真时设为1，综合时设为0
```

验收与提交

- 检查Block Design电路模块图 (2分)
- 检查实验现象并回答问题 (4分)
- 将源码、运行结果、实验报告打包提交 (6分)
 - 命名规则: **学号_姓名_DLA实验3.zip**
 - 提交方法: <https://hitsz-cslab.gitee.io/dla/ojguide>
 - DDL: 下周同一上课时间前
- 附加题 (+1~3.5分):
 - 检查实验现象、回答问题并提交源码、运行结果和报告



哈爾濱工業大學(深圳)

HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

开始实验