法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

第9课　图像描述（图说）

Image Captioning

主讲人：张宗健

悉尼科技大学博士

主要研究方向：　计算机视觉、视觉场景理解、图像&语言、深度学习

图像检索CbIR、Human ReID等

本章结构

- 深度语言模型介绍
- LSTM原理解析
- 图说模型原理与结构
- 大数据集介绍
- 应用实例：开源模型Show and Tell

深度语言模型

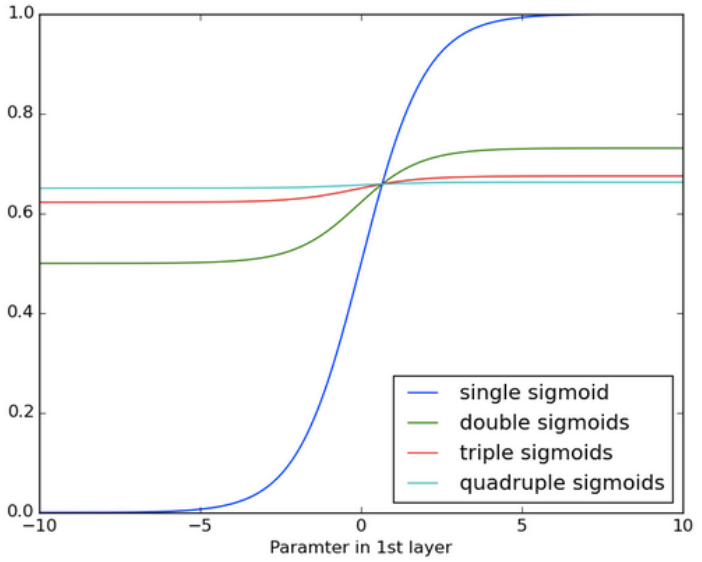
递归神经网络RNN

- 有2类
  - 时间递归神经网络(Recurrent Neural Network)
    - 针对时间序列
  - 结构递归神经网络(Recursive Neural Network)
    - 针对树状结构
- 优化方法
  - 时序后向传播(Back propagation through time)
- 长时记忆/递归深度问题
  - 梯度爆炸(Gradient exploding) → 梯度剪切
  - 梯度消失(Gradient vanishing) → 特殊设计

深度语言模型

时序后向传播（BPTT）

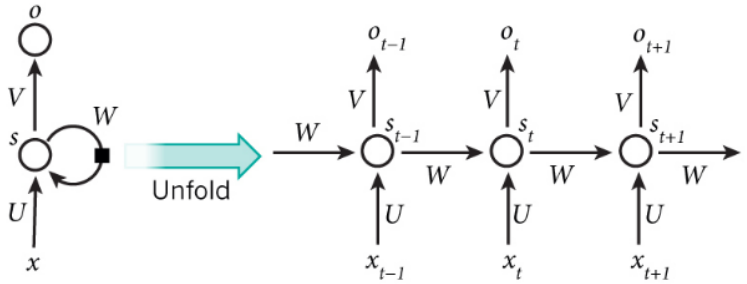
- 传统后向传播(BP)在时间序列上的扩展
- t时刻的梯度是前t-1时刻所有梯度的累积
- 时间越长，梯度消失越严重

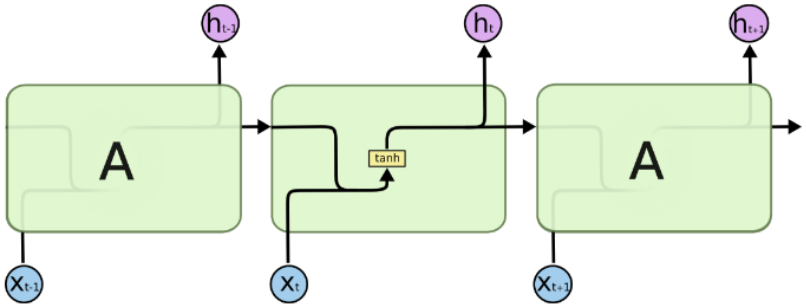


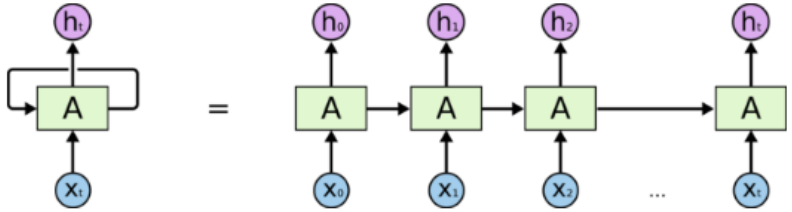
深度语言模型

朴素Vanilla-RNN

- 单层神经网络在时间上的扩展
- t-1时刻的隐层状态(Hidden state)会参与t时刻输出的计算
- 严重的梯度消失问题







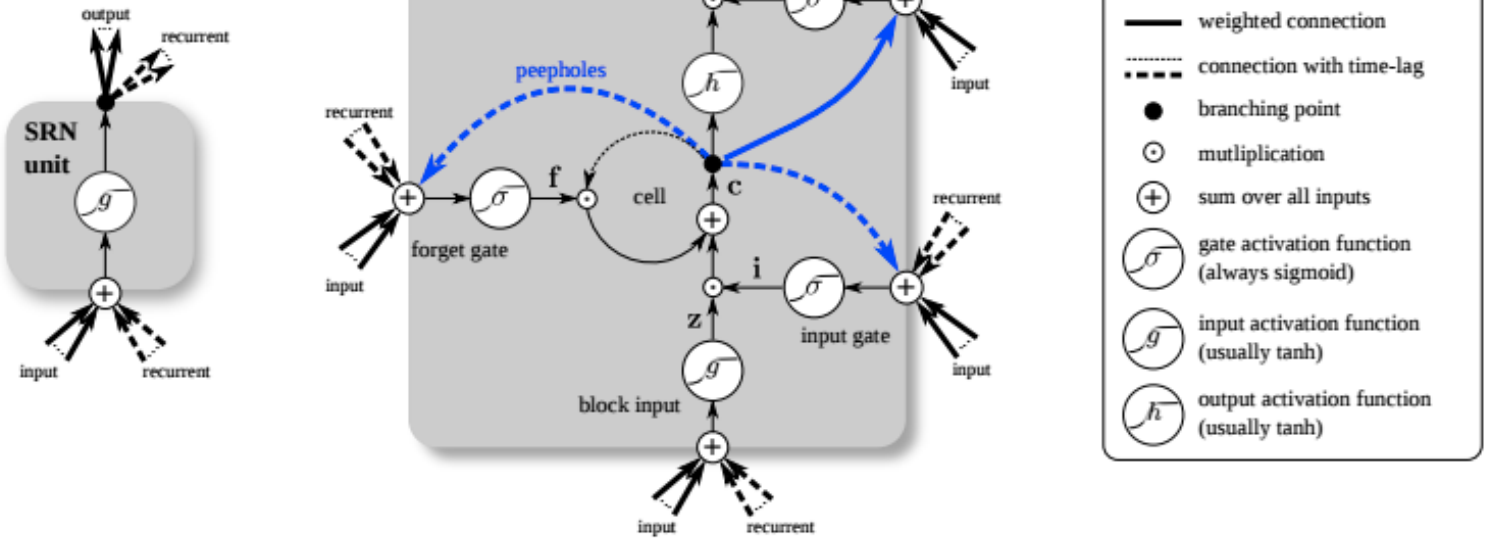
深度语言模型

LSTM长短时记忆模型(Long Short-Term Memory)

- Hochreiter & Schmidhuber 于1997提出
- 有效捕捉长时记忆(Long dependency)
- 包含4个神经元组
  - 1个记忆神经元(Memory cell)
  - 3个控制门神经元
    - 输入门(Input gate)
    - 忘记门(Forget gate)
    - 输出门(Output gate)

深度语言模型

Vanilla-RNN vs LSTM



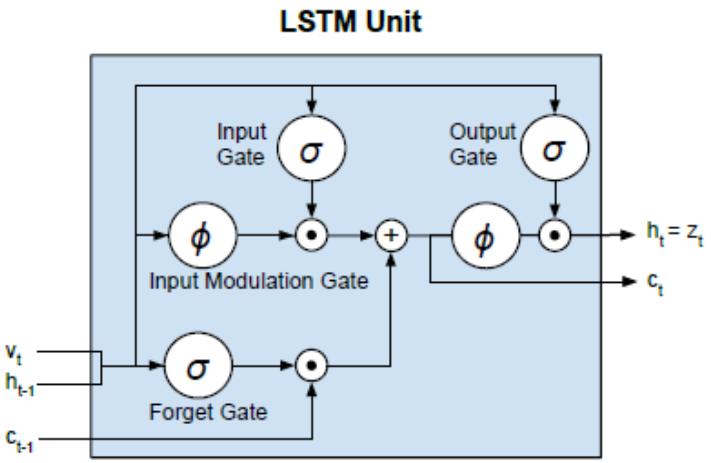
Vanilla-RNN

LSTM

深度语言模型

LSTM数学模型

- 3个输入
  - 前一时刻的隐含状态h_{t-1}
  - 前一时刻的记忆状态c_{t-1}
  - 当前时刻的输入x_t
- 2个输出
  - 当前时刻的隐含状态h_t
  - 当前时刻的记忆状态c_t



$$i_t = \sigma(W_{vi}v_t + W_{hi}h_{t-1} + b_i) \quad (3)$$

$$f_t = \sigma(W_{vf}v_t + W_{hf}h_{t-1} + b_f) \quad (4)$$

$$o_t = \sigma(W_{vo}v_t + W_{ho}h_{t-1} + b_o) \quad (5)$$

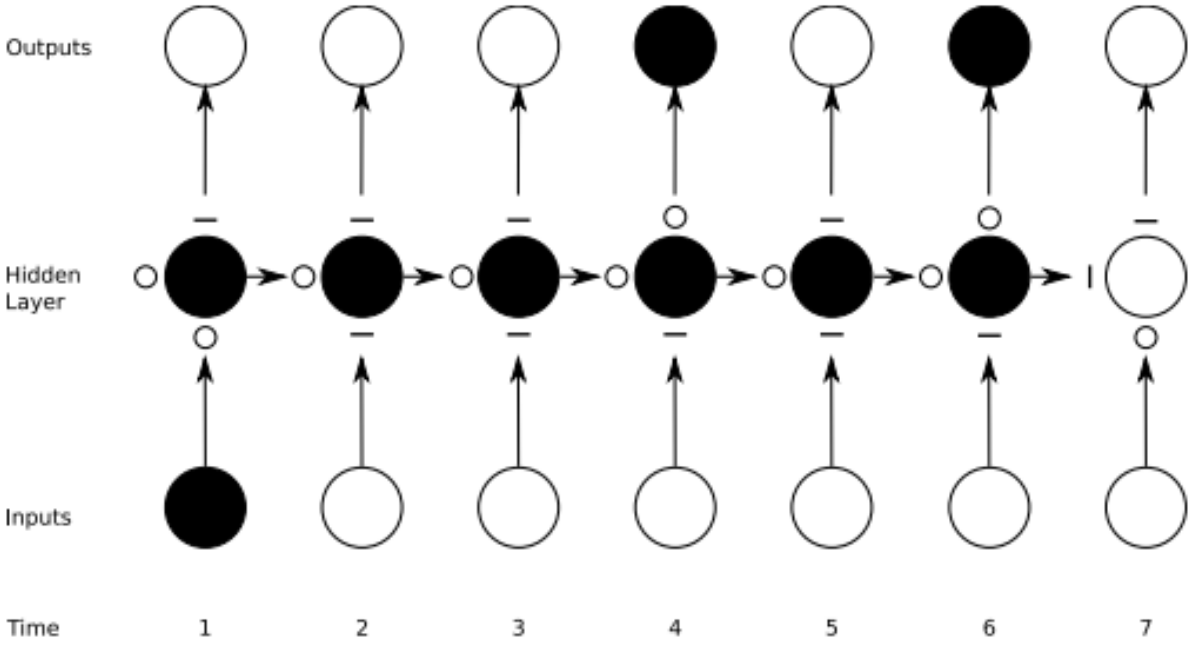
$$g_t = \phi(W_{vg}v_t + W_{hg}h_{t-1} + b_g) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (7)$$

$$h_t = o_t \odot \phi(c_t) \quad (8)$$

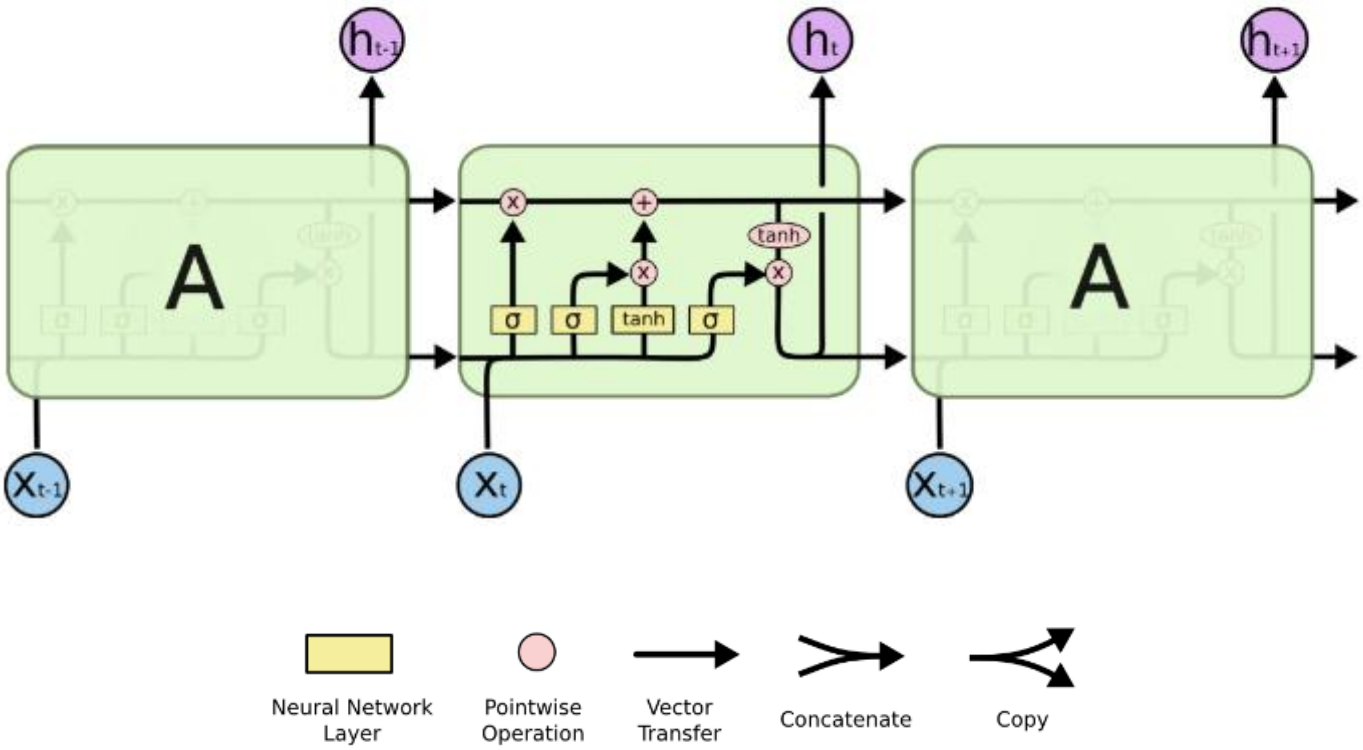
深度语言模型

LSTM控制门作用



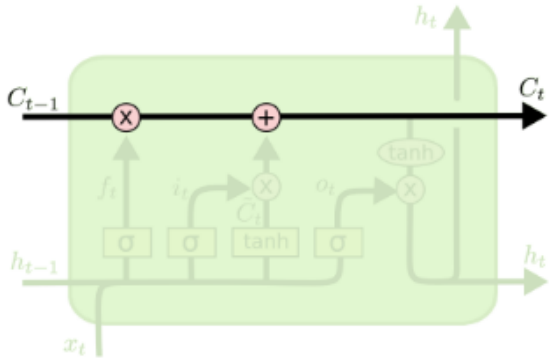
深度语言模型

LSTM结构图

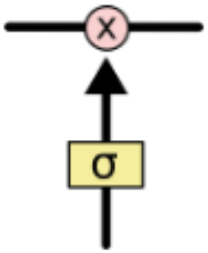


深度语言模型

LSTM结构图



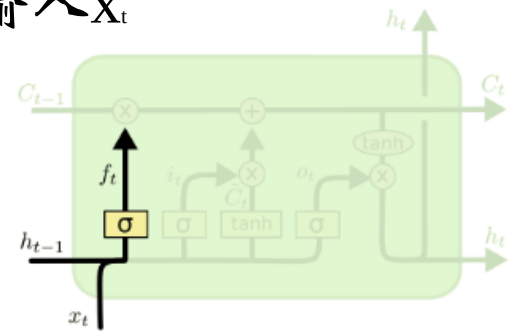
- 记忆状态(cell state) → 信息
  - 存储之前时刻的信息
  - 避免长时记忆问题的核心
- 控制门(gate) → 选择性控制信息流入
  - 由元素乘操作实现
  - 配有sigmoid激活函数的神经层
  - 值域[0,1]
    - 0 不通过任何信息
    - 1 通过所有信息



深度语言模型

LSTM结构图

- 忘记门
  - 决定前一时刻中多少记忆状态被移除
  - Sigmoid激活
  - 2个输入
    - 前一时刻的隐含状态h_{t-1}
    - 当前时刻的输入x_t

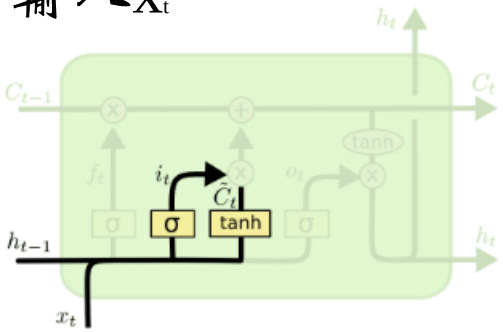


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

深度语言模型

LSTM结构图

- 输入门
  - 决定当前时刻有多少新输入信息需要存入记忆状态
  - Sigmoid激活
  - 2个输入
    - 前一时刻的隐含状态h_{t-1}
    - 当前时刻的输入x_t



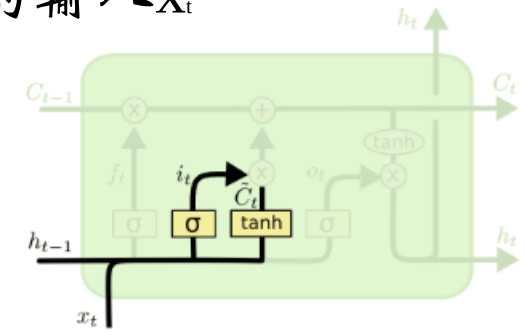
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

深度语言模型

LSTM结构图

- 输入调制
  - 决定当前时刻有多少新输入信息需要存入记忆状态
  - Tanh激活
  - 2个输入
    - 前一时刻的隐含状态h_{t-1}
    - 当前时刻的输入x_t



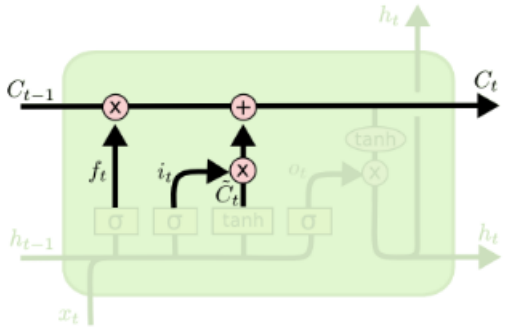
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

深度语言模型

LSTM结构图

- 记忆状态更新
  - 选择性移除前一时刻态的旧信息（记忆状态）
  - 选择性添加当前时刻的新信息（调制输入）

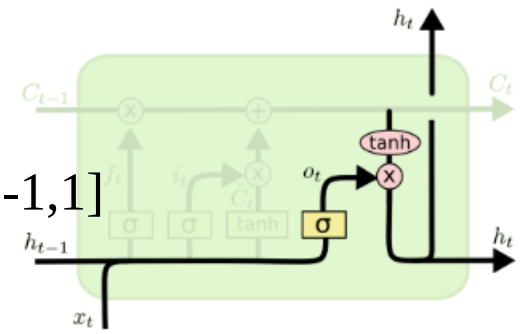


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

深度语言模型

LSTM结构图

- 输出门
  - 决定当前时刻多少记忆状态用于输出
  - 2个输入
    - 前一时刻的隐含状态h_{t-1}
    - 当前时刻的输入x_t
  - 2个激活
    - Tanh激活
      - 压缩记忆状态[-1,1]
    - Sigmoid激活

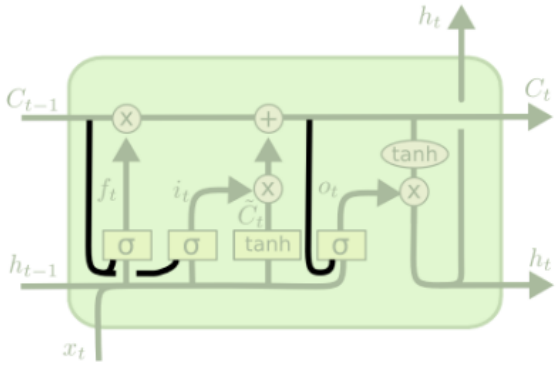


$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

深度语言模型

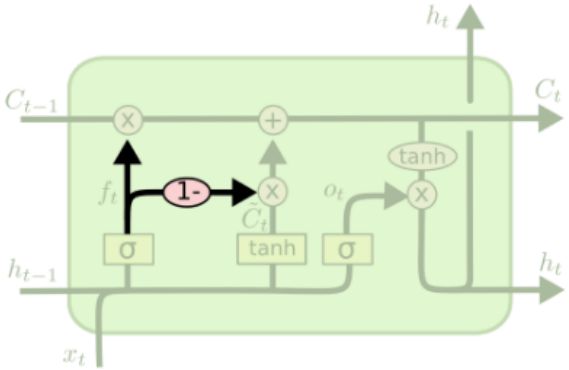
LSTM变种

- Peephole



$$f_t = \sigma\left(W_f \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [\boldsymbol{C_{t-1}}, h_{t-1}, x_t] + b_i\right)$$
$$o_t = \sigma\left(W_o \cdot [\boldsymbol{C_t}, h_{t-1}, x_t] + b_o\right)$$

- Coupled 忘记-输入门

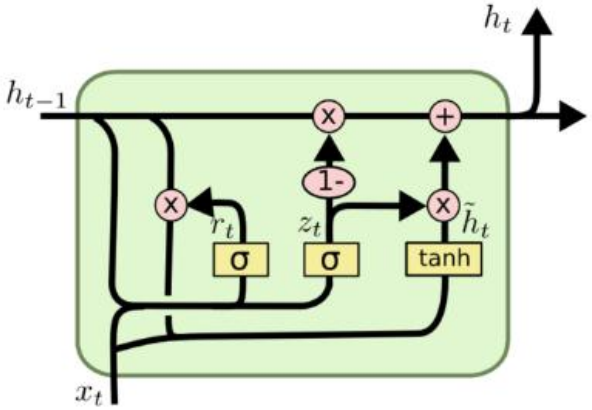


$$C_t = f_t * C_{t-1} + (1 - \boldsymbol{f_t}) * \tilde{C}_t$$

深度语言模型

GRU门限递归单元（Gated Recurrent Unit）

- 2个改动（Cho, et al., 2014）
  - 合并输入门和忘记门
  - 合并记忆状态和隐藏状态
- 2个控制门
  - 重置门(Reset gate)
  - 更新门(Update gate)
- 2个输入（与LSTM一致）
- 1个输出



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

深度语言模型

LSTM vs GRU

- LSTM
  - 模型复杂，参数多，拟合能力强
  - 数据要求：大规模、复杂度高
- GRU
  - 模型精简，参数少，拟合能力相对弱
  - 适用于偏小规模、不是很复杂的数据

图说模型Image Captioning

为图片生成描述语言

- 输入：图片
- 输出：客观描述图片内容的句子

A person is holding a **gecko** in their hand.

A large **blimp** in a blue sky.

A black and white **skunk** is eating grass.

A plate of food with a fork and a **hollandaise**.

图说模型Image Captioning

一个视觉-语言研究问题

- 理解为一种特殊的机器翻译：视觉→语言
- 模型需要有复杂的场景理解能力
  - 图片理解→计算机视觉(Computer Vision)
  - 语言理解→自然语言处理(Natrual Language Processing)
  - 复合、多模态理解→多媒体(Multi-Media)

图说模型Image Captioning

研究难点与挑战

- 多模态理解与推理
  - 图片：捕捉真实世界的原始刻画
  - 自然语言：代表更高一级的抽象
- 复合理解与推理
  - 多个元素：物体、动作、场景、事件等
  - 多步、迭代过程

图说模型Image Captioning

理解模式

1. 完整理解图片所有内容
2. 用语言描述出自己的理解

模型策略

1. 传统的分段处理策略
2. 新的点对点策略(End-to-end trainable way)

图说模型Image Captioning

传统的分段处理策略

- 流程
  1. 图片内容→文本标签
  2. 文本标签→描述语句
- 优势
  - 虑除干扰信息
  - 模块化结构
  - 直接使用CV和NLP的研究成果
- 劣势
  - 第一步中错误判断会单向影响第二步的语言推理

图说模型Image Captioning

传统的分段处理策略

- 流程
  1. 将图片跟文本映射到同一共享空间下
  2. 翻译：图片特征→语言描述
- 优势
  - 同时训练，最优协作
  - 模块化结构
- 劣势
  - 黑箱严重

图说模型Image Captioning

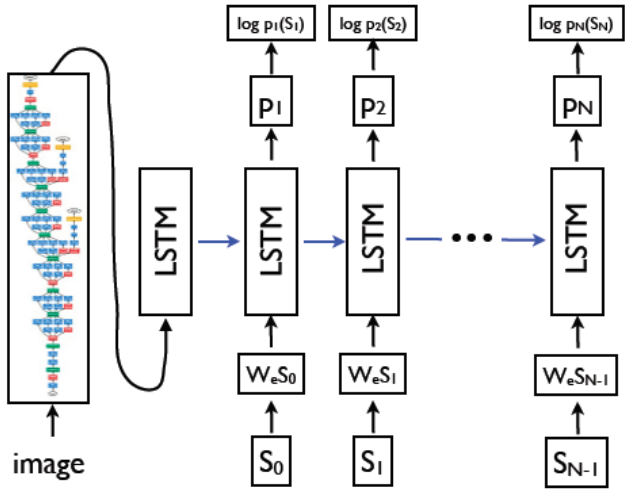
State-of-the-art模型组成

- DNN框架
  - CNN: 图片理解
    - VGG, ResNet
  - RNN: 语言理解及生成
    - Multimodal-RNN, LSTM, GRU
  - 特殊功能模块
    - Attention

图说模型Image Captioning

Show and tell模型

- CNN + LSTM
- CNN: Inception v3生成图片特征（最后全连接层）
- 特征映射矩阵We: 将文本映射到图片特征空间
- 文本编码：one-hot
- LSTM
  - CNN特征作为第一个词
  - 句子中的词作为后续序列



图说模型Image Captioning

Show and tell模型

- LSTM语言生成器

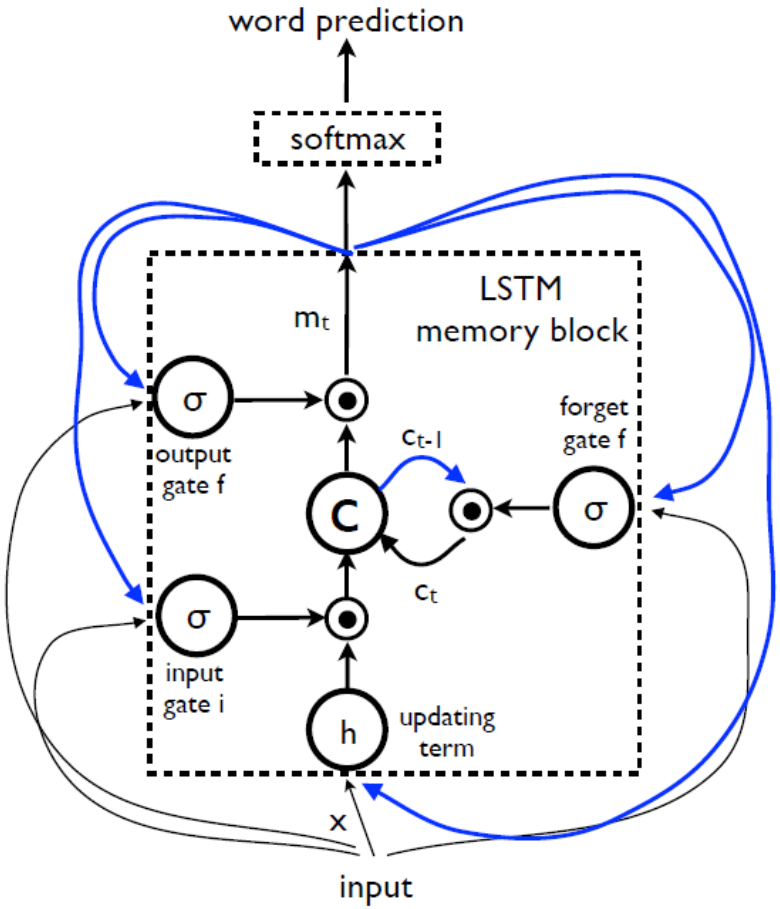
$$
\begin{aligned}
i_t &= \sigma(W_{ix}x_t + W_{im}m_{t-1}) \\
f_t &= \sigma(W_{fx}x_t + W_{fm}m_{t-1}) \\
o_t &= \sigma(W_{ox}x_t + W_{om}m_{t-1}) \\
c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}m_{t-1}) \\
m_t &= o_t \odot c_t \\
p_{t+1} &= \text{Softmax}(m_t)
\end{aligned}
$$

- CNN+LSTM图说

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

$$
\begin{aligned}
x_{-1} &= \text{CNN}(I) \\
x_t &= W_e S_t, \quad t \in \{0 \ldots N-1\} \\
p_{t+1} &= \text{LSTM}(x_t), \quad t \in \{0 \ldots N-1\}
\end{aligned}
$$

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$



图说模型Image Captioning

Show and tell模型

- 训练细节
  - 第一步：固定CNN参数，训练LSTM语言模型500K
    - CNN参数：在ImageNet数据集预训练(pre-trained)好的参数
    - 训练拆分：一句话n个词→n-1组训练序列
  - 第二步：细调CNN参数，CNN&LSTM一起训练100K
- 推理策略
  - Beam Search (尺寸=3)
  - 每一步获取Top3概率的词作为备选

图说模型Image Captioning

Show, attend and tell(SAT)模型

- CNN + LSTM + Attention module（注意机制）
- CNN: VGG生成图片特征
- 特征映射矩阵We: 将文本映射到图片特征空间
- 文本编码：one-hot
- LSTM
  - 添加第3输入：基于attention的图片特征

图说模型Image Captioning

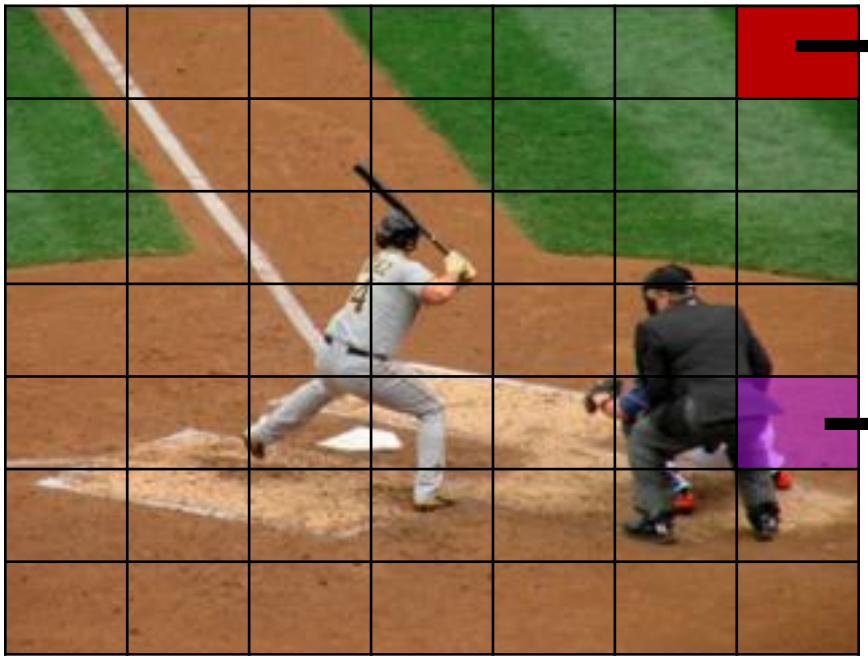
Show, attend and tell(SAT)模型

- CNN + LSTM + Attention module（注意机制）
- CNN
  - VGG最后卷基层生成图片特征
- 特征映射矩阵We: 将文本映射到图片特征空间
- 文本编码
  - one-hot
- LSTM
  - 添加第3输入：基于attention的图片特征

图说模型Image Captioning

注意机制的CNN特征

- VGG最后卷基层输出14x14x512



14 X 14 网格拆分

1	2	…	511	512	
					1
					2
					…
					14
					…
					…
					…
					…
					140
					…
					…
					196

图说模型Image Captioning

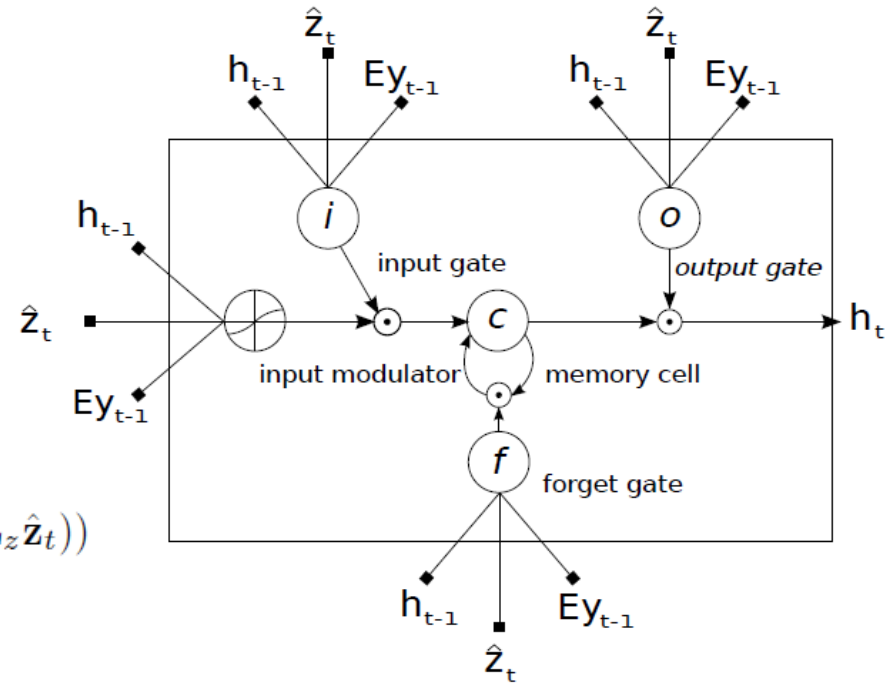
SAT模型 Attention module（注意机制）

- LSTM改进

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$



- 输出推断

$$p(\mathbf{y}_t|\mathbf{a}, \mathbf{y}_1^{t-1}) \propto \exp(\mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h\mathbf{h}_t + \mathbf{L}_z\hat{\mathbf{z}}_t))$$

- 初始化
  - 特征均值

$$\mathbf{c}_0 = f_{\text{init,c}}(\frac{1}{L}\sum_i^L \mathbf{a}_i)$$

$$\mathbf{h}_0 = f_{\text{init,h}}(\frac{1}{L}\sum_i^L \mathbf{a}_i)$$

图说模型Image Captioning

SAT模型 Attention module（注意机制）

- 注意机制模块
  - 注意权重推断

$$e_t = f_{att}(a, h_{t-1}) = W_{att}\tanh(W_a a + W_h h_{t-1} + b)$$

$$\alpha_t = \text{softmax}(e_t)$$

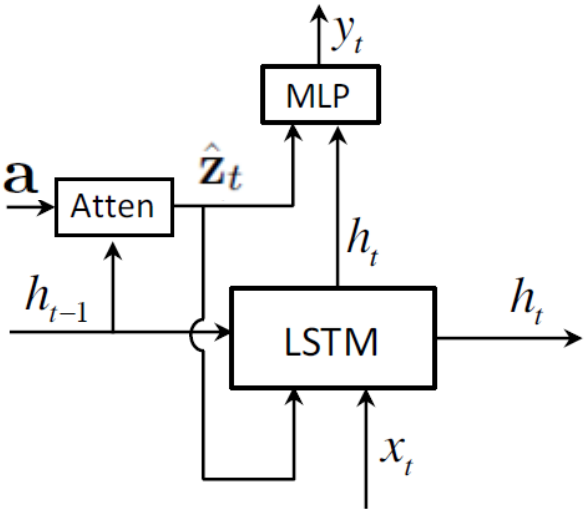
    - 14x14特征中单个特征

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L}\exp(e_{tk})}$$

  - 特征融合（权重相加）

$$\hat{\mathbf{z}}_t = \phi(\{\mathbf{a}_i\}, \{\alpha_i\}) = \sum_i^L \alpha_i \mathbf{a}_i = \alpha_t \cdot \mathrm{a}$$



图说模型Image Captioning

SAT模型 Attention module（注意机制）

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

图说模型Image Captioning

数据集

- MSCOCO标注集&竞赛
  - 描述(Captions)
  - 80个类别的物体(Object)语义分割
  - 100,000人的肢体关键点(Keypoints)
  - 其他附属标注集

	训练集	评估集	测试集	图片标注
MSCOCO	82783	40504	40775	5句描述
Flickr30k	28000	1000	1000	5句描述
Flickr8k	6000	1000	1000	5句描述

图说模型Image Captioning

性能指标

- METEOR
  - 与人类评判结果最相关（接近）
- CIDER
  - 与人类评判结果次相关
- BLEU@N（N代表n-gram，分别是1,2,3,4）
- ROUGE-L
- Perplexity

演示环节

- Github
  - https://github.com/349zzjau
- 百度网盘
  - http://pan.baidu.com/s/1gfpCCwj
- 代码演示
  - Show and tell

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

Q & A

小象账号：349zzjau

课程名：基于深度学习的计算机视觉

课后调查问卷http://cn.mikecrm.com/U9euAYY

Reference List

[1] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2017). Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, *39*(4), 652-663.

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057).

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象
- 新浪微博：ChinaHadoop