lecture04 Backpropagation and computation graphs

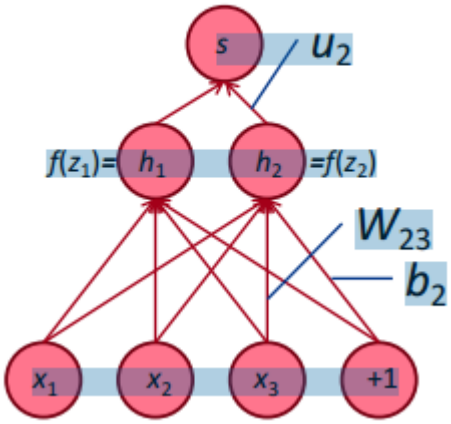
Derivative wrt a weight matrix

继续计算∂s/∂W

$$\frac{\partial s}{\partial W} = \delta \frac{\partial z}{\partial W} = \delta \frac{\partial}{\partial W} Wx + b$$

先考虑单个权重Wij的导数：
Wij只对 zi 有贡献。例如W23只对z2有贡献，对z1没有贡献

$$\frac{\partial zi}{\partial Wij} = \frac{\partial}{\partial W_{ij}} W_i \cdot x + b_i = \frac{\partial}{\partial W_{ij}} \sum_{k=1}^{d} W_{ik} x_k = x_j$$



对于单个Wij:

$$\frac{\partial s}{\partial W_{ij}} = \delta_i x_j$$

对于整体权重W的梯度，有：

$$\frac{\partial s}{\partial W} = \delta^T x^T$$

其中δ为(n*1)列向量，x为(1* m)行向量

梯度推导tips：

- 定义变量并注意他们的维度

- 求导链式法则
- 不同条件下分别考虑
- 逐个元素求偏导
- 检查维度

Deriving gradients wrt words for window model

更新词向量的梯度可以被简单的分为每个词向量的梯度

$$\delta_{window} = \begin{bmatrix} \nabla x_{museums} \\ \nabla x_{in} \\ \nabla x_{paris} \\ \nabla x_{are} \\ \nabla x_{amazing} \end{bmatrix} \in \mathrm{R}^{5d}$$

逐个更新词向量有助于命名实体的分类，但是更新非常稀疏

A pitfall when retraining word vectors

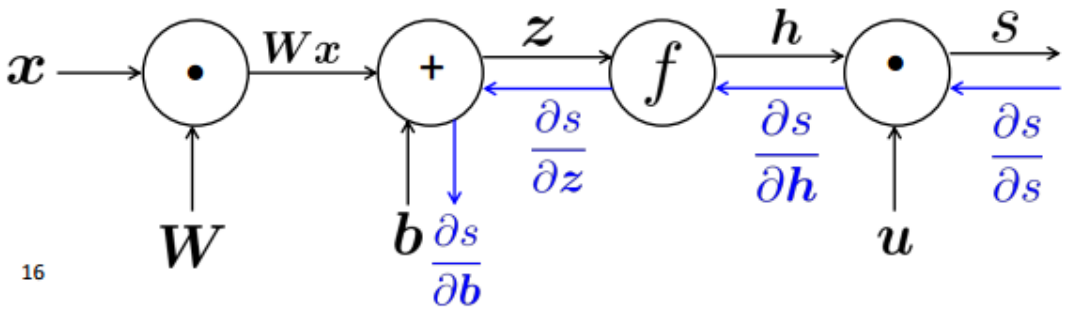
同义的训练数据与测试数据，在情感分类中训练数据在更新过程中会不断变化，而测试数据不会

So what should I do?

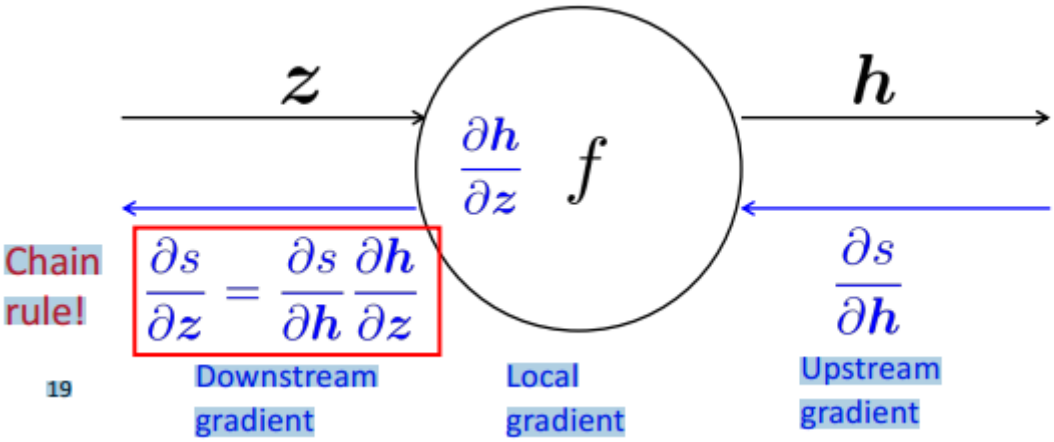
1.是否应该预训练词向量？。是，因为：训练算法很简单；只需在大量数据上进行
2.当训练有监督分类器时，是否应该更新词向量？训练数据集很小的时候不要训练词向量；数据集很大的时候最好训练-更新-调整词向量。

Backpropagation

trick：在计算低层导数时，重用高层的导数，可以减少计算量。
computation graph（前向传播）：fbe85997b9b4253bb008f05d73c1e7be.png](en-resource://database/1922:0)
反向传播：沿着边反向传递梯度



以单个节点为例：同样是使用链式准则



当有多个输入时：