后厂理工学院
houchangtech.com

HCT NLP Week 4

问答摘要与推理
Seq2Seq（二）

Outline

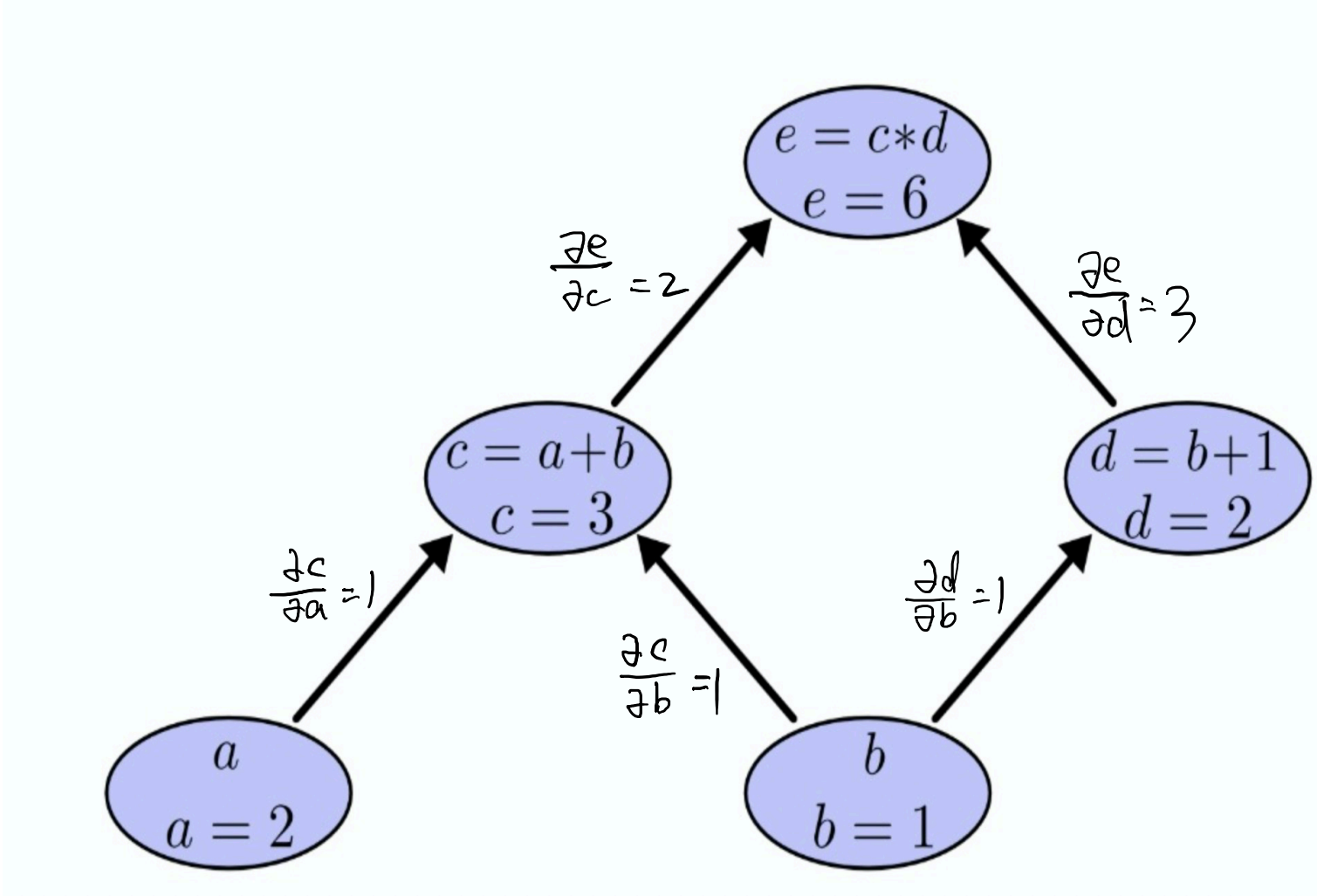
- 深度学习框架图计算理论
- Beam search
- 生成式文本摘要问题补充
- Baseline代码实践

Outline

- 深度学习框架图计算理论
- Beam search
- 生成式文本摘要问题补充
- Baseline代码实践

图计算理论

Computational Graphs



$e = (a+b) * (b+1)$

$c = a+b$

$d = b+1$

$e = c * d$

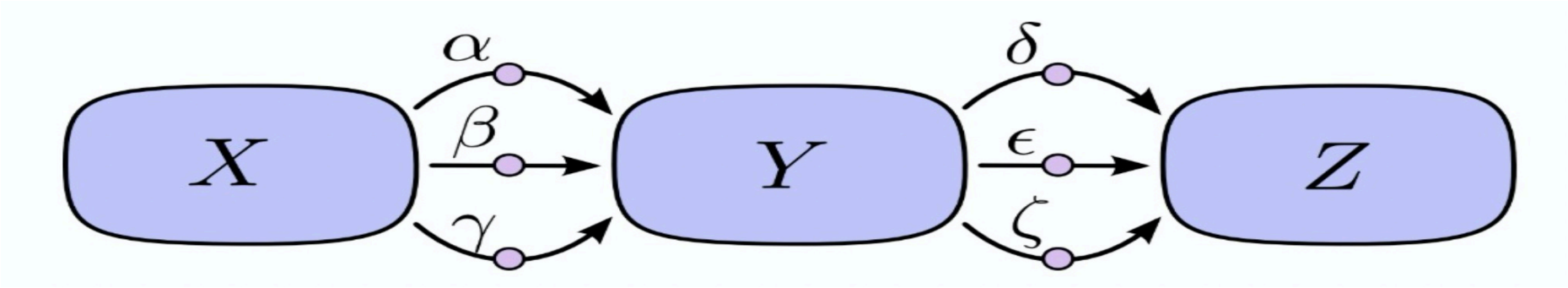
$\frac{\partial e}{\partial b}$ Derivatives on it ?

$\frac{\partial e}{\partial b} = 1 \times 2 + 1 \times 3 = 5$

图计算理论

Factoring Paths

$$\frac{\partial z}{\partial x} = \alpha\delta + \alpha\epsilon + \alpha\zeta + \beta\delta + \beta\epsilon + \beta\zeta + \gamma\delta + \gamma\epsilon + \gamma\zeta = (\alpha+\beta+\gamma)(\delta+\epsilon+\zeta)$$

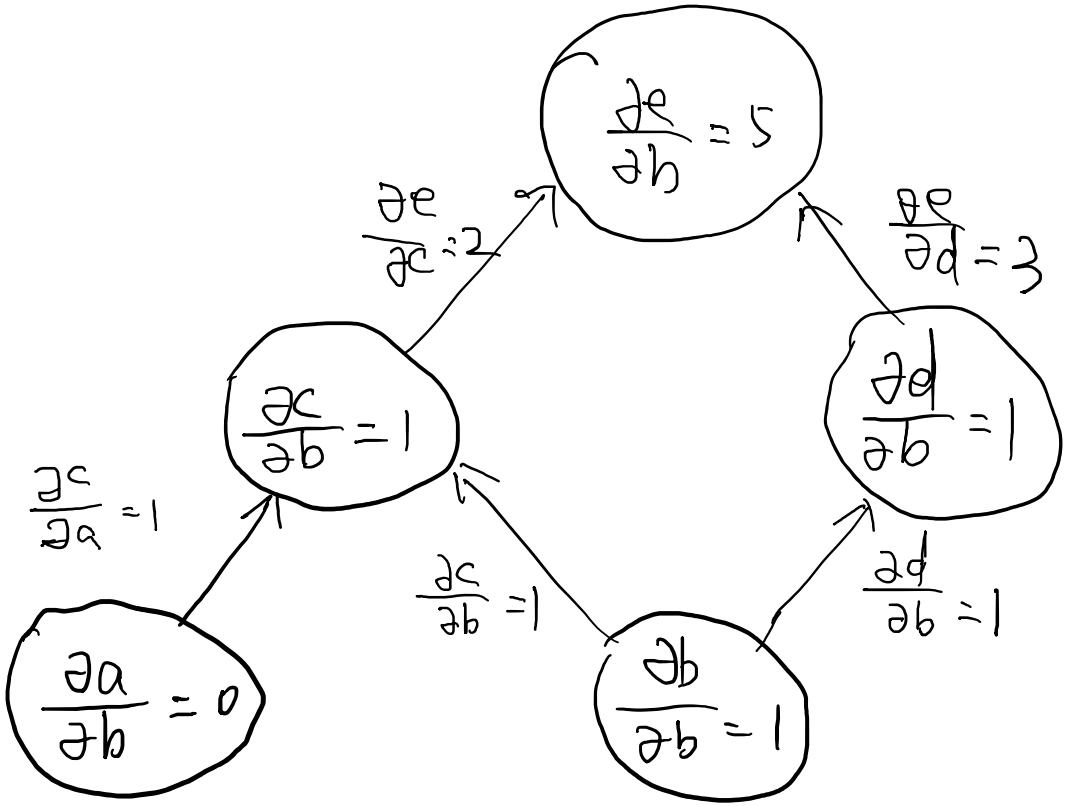


forward-mode differentiation

reverse-mode differentiation

图计算理论

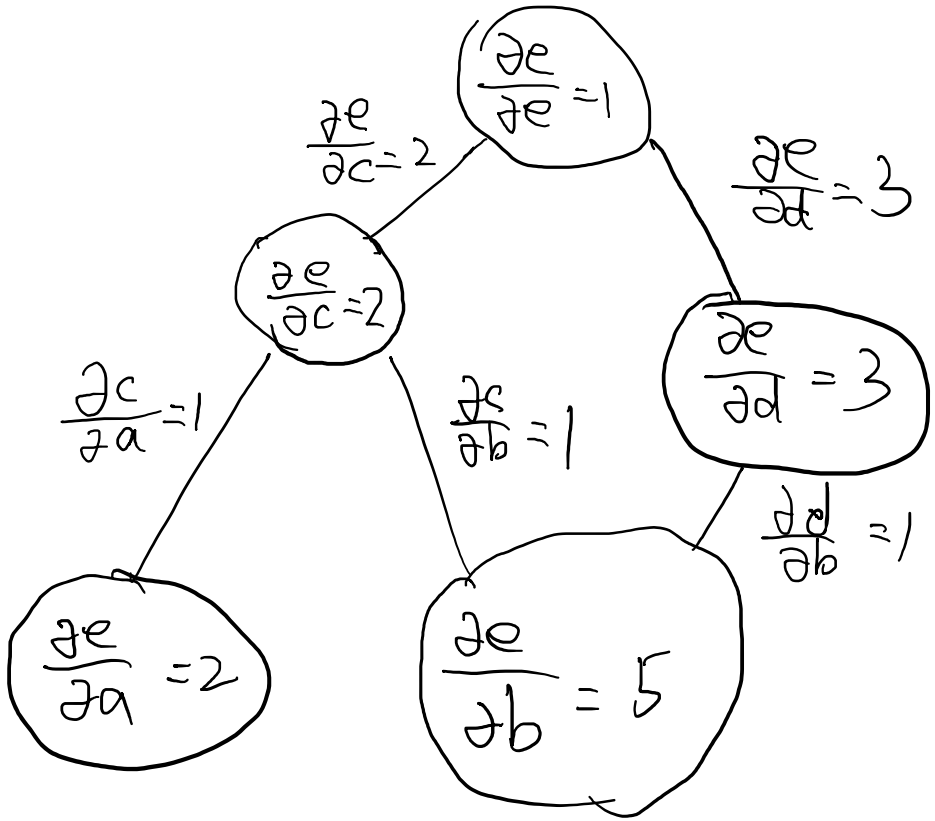
forward-mode differentiation



这张图从下从上指

图计算理论

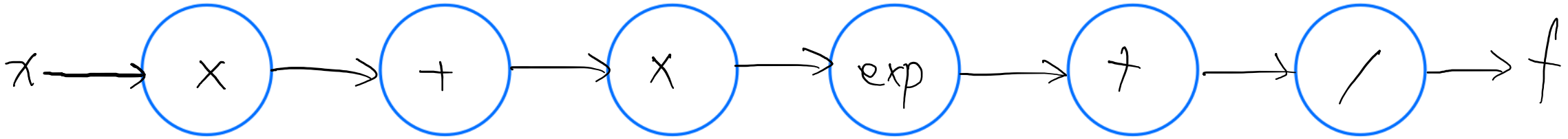
reverse-mode differentiation



这张图从上往下推

图计算理论

$$f(x;w,b)=\frac{1}{\exp\big(-(wx+b)\big)+1}$$



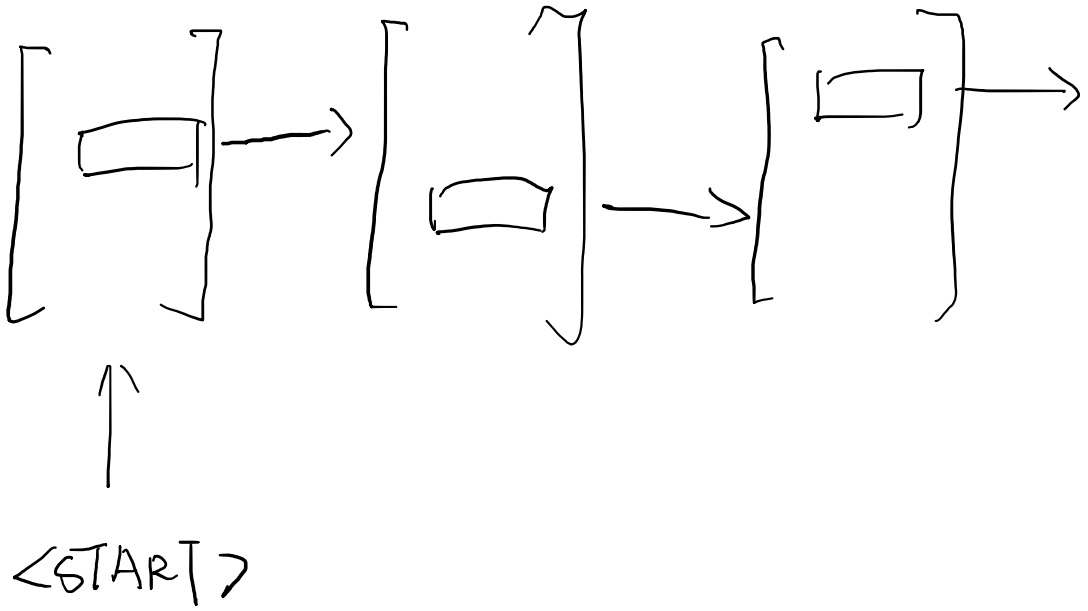
Outline

- 深度学习框架图计算理论
- Beam search
- 生成式文本摘要问题补充
- Baseline代码实践

Beam search

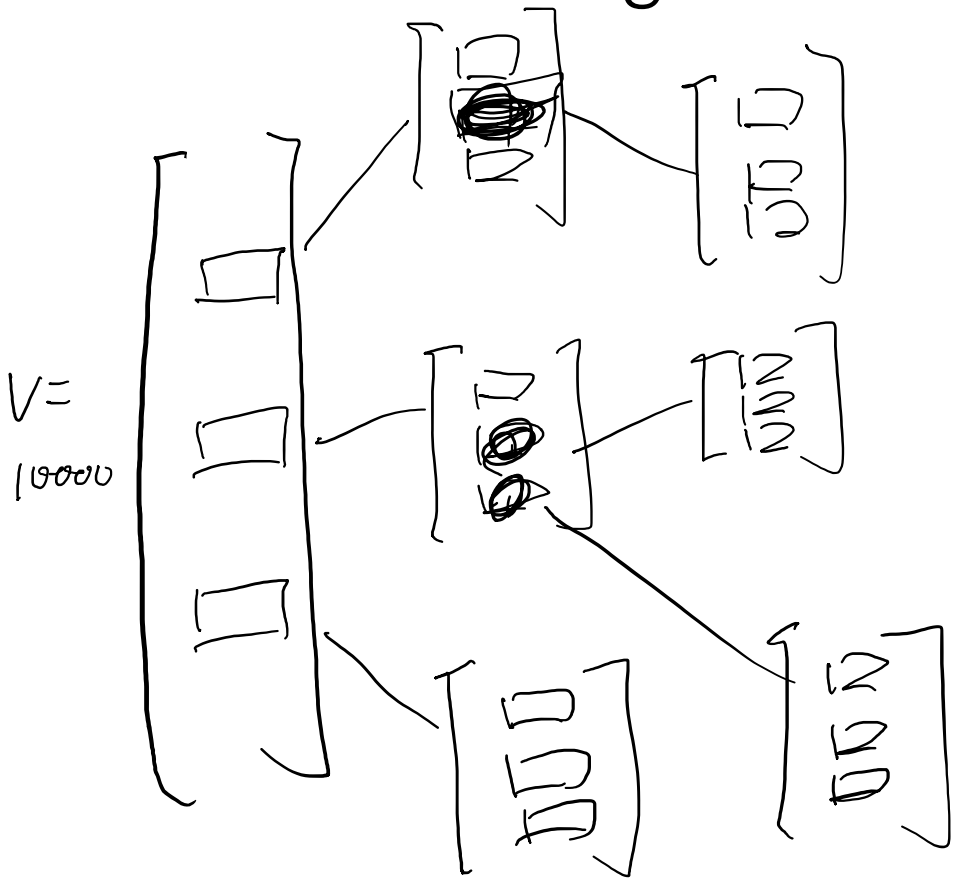
greedy search algorithm

行波 没有 极限 的 感觉



Beam search

Beam search algorithm

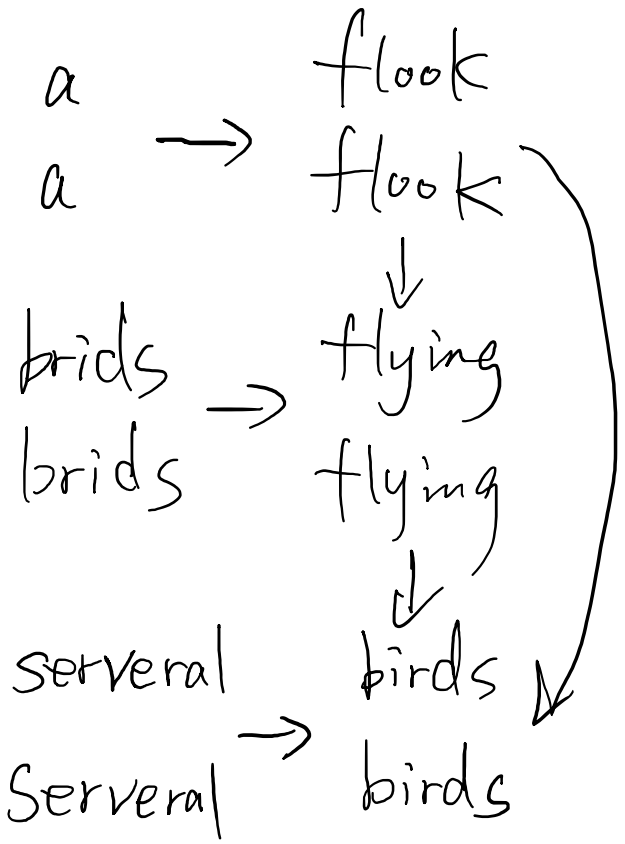


Beam search

Beam search algorithm

Beam search

Beam search algorithm



B=6，G=3，每一组的 beam width 为 2

打分 scores

此时刻可能输出的词在上面的组别现过，我们对这个词分数减 1，没出现，不减罚。

diverse beam search

SUMMARY GENERATION

Diverse Beam Decoding

the top-B hypotheses may differ by just a couple tokens at the end of sequences, which not only affects the quality of generated sequences but also wastes computational resources

Outline

- 深度学习框架图计算理论
- Beam search
- 生成式文本摘要问题补充
- Baseline代码实践

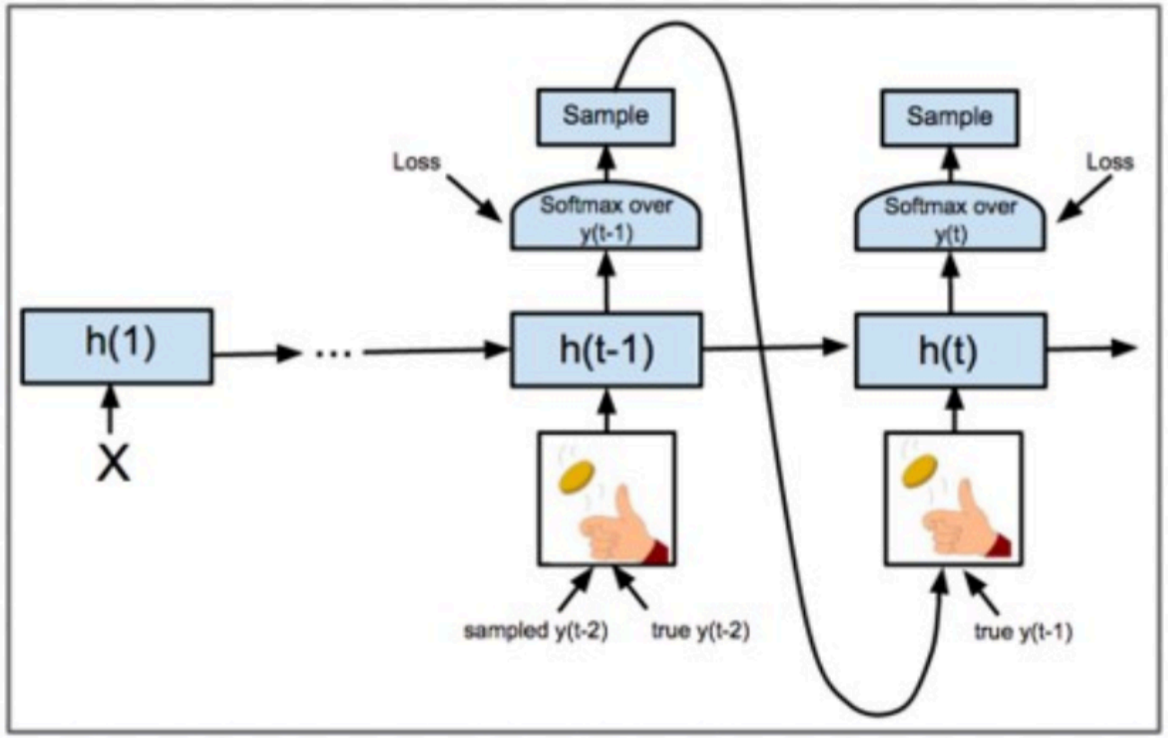
Scheduled Sampling

A method for avoiding the problem of exposure bias.

是一种解决训练和生成时输入数据分布不一致的方法。在训练早期该方法主要使用目标序列中的真实元素作为解码器输入，可以将模型从随机初始化的状态快速引导至一个合理的状态。随着训练的进行，该方法会逐渐更多地使用生成的元素作为解码器输入，以解决数据分布不一致的问题。该方法应用在模型的训练阶段，生成阶段不使用。

Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks

Scheduled Sampling



定义 $\varepsilon \in [0, 1]$

$\varepsilon \cdot X_{t-1} + (1-\varepsilon)\hat{X}_{t-1}$

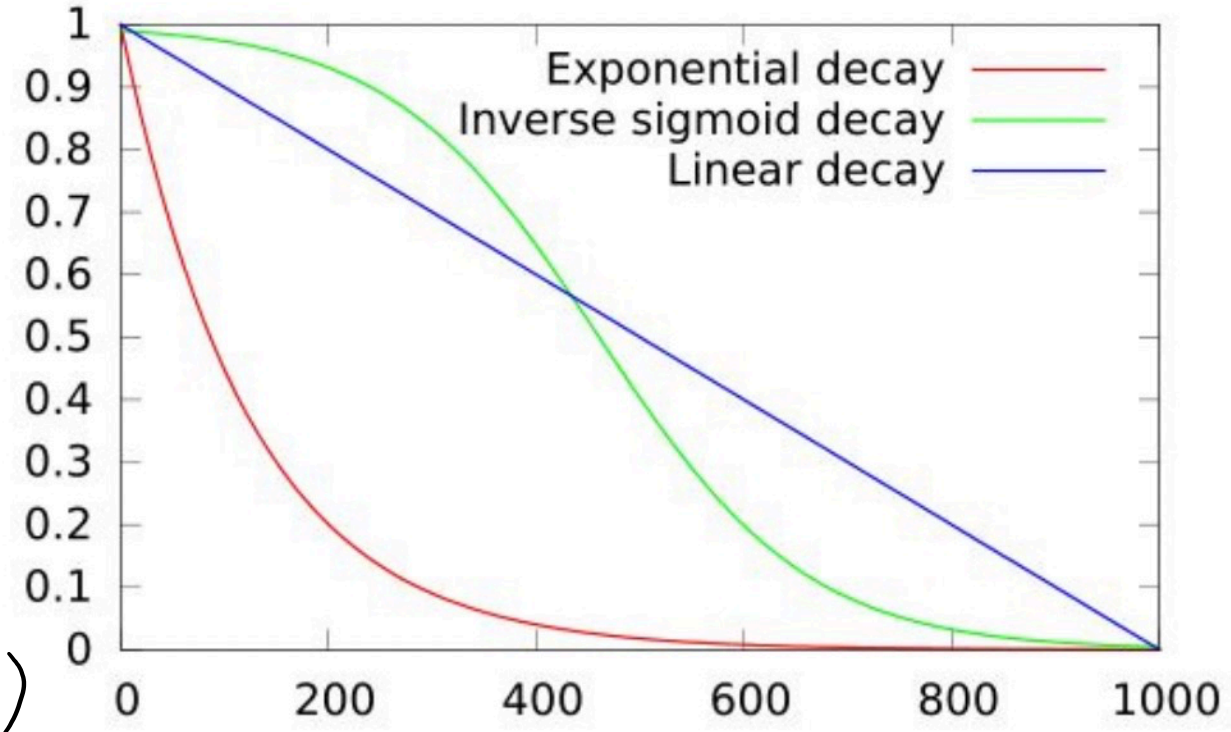
真实　　预测

(1) 线性 decay

$\varepsilon_i = \max(\varepsilon, k - ci)$

(2) 指数 decay　　$\varepsilon_i = k^i$

(3) 逆 sigmoid decay　　$\varepsilon_i = k/(k+\exp(i/k))$



Datasets

CNN dailymail数据集

First highlight: Argentina coach Sabella believes Messi's habit of being sick during games is down to nerves.

First 2 sentences: Argentina coach Alejandro Sabella believes Lionel Messi's habit of throwing up during games is because of nerves. The Barcelona star has vomited on the pitch during several games over the last few seasons and appeared to once again during Argentina's last warm-up match against Slovenia on Saturday.

单文档摘要：
Gigaword
LCSTS
Newsroom
Xsum

新浪微博摘要数据集（679898 条数据）

【"干杯 大哥！"外卖小哥点头那一刻 泪目😭】近日，一男生和外卖小哥之间的互动，在网上刷屏。视频中，外卖小哥坐在电瓶车上啃干粮，男生请他帮拧瓶盖。当外卖小哥拧开第二瓶水时，男生说"干杯，大哥，天气很热，加油！"。外卖小哥这才反应过来，点头致谢❤️ （央视）

train_text.txt	短文本的内容，约100-200字
train_label.txt	短文本的摘要，约10-20字

ROUGE

recall-oriented understand for gisting evalution

- ROUGE-N
- ROUGE-L
- ROUGE-S
- ROUGE-W
- ROUGE-SU

$$Rouge\_N = \frac{\text{参考摘要和自动摘要共有 n-gram 个数}}{\text{n-gram 的个数（参考摘要）}}$$

自动摘要 the cat was found under the bed (Y)

参考摘要 the cat was under the bed (X_1)

$$Rouge\_1(X_1, Y) = \frac{6}{6} = 1.0$$

$$Rouge\_2(X_1, Y) = \frac{4}{5} = 0.8$$

ROUGE

recall-oriented understand for gisting evalution

- ROUGE-N
- ROUGE-L
- ROUGE-S
- ROUGE-W
- ROUGE-SU

police ended the human 3/4

police killed the gunman

the gunman murdered police 2/4

$R_{lcs} = \frac{LCS(X,Y)}{m}$ m为参考摘要长度

$P_{lcs} = \frac{LCS(X,Y)}{n}$ n为候选摘要长度

$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$

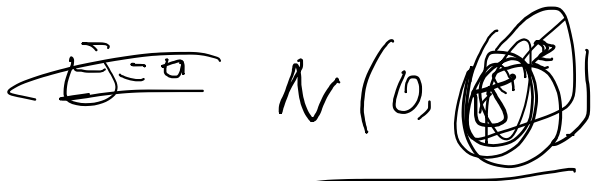
→ 取 $\beta = 1$

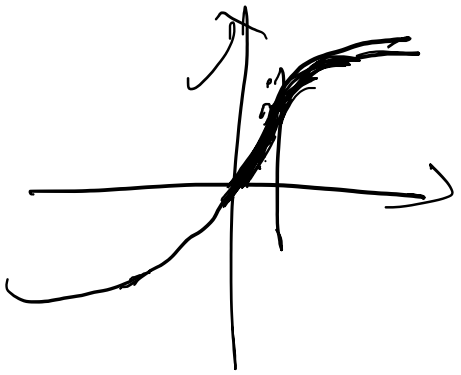
$Rouge\_L = \frac{2}{4}$

S_2 优于 S_3

Initializing neural networks

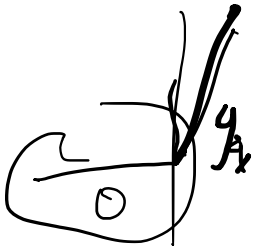
(1) 固定方差 < 高斯分布的分布





(2) 基于方差缩放

Xavier初始化 (Glorot) → Sigmoid tanh



He初始化 → Relu

Relu

Batch size

1. Batch size是用于在每次迭代中训练模型的数据数量。一般的设置是32，64，128，256，512。

2. 选择正确的Batch size对于确保cost function和参数值的收敛，以及模型的泛化能力。

3. Batch size决定更新的频率。Batch size越小，更新就越快。

4. Batch size越大，梯度越精确。也就是说，在迭代计算的时候更容易跳过局部区域。

5. 比较大Batch size，往往GPU memory是不够用的，就需要通过并行计算的方式解决。

Choice of optimizer

(Stochastic) Gradient Descent

Momentum

RMSprop

Adam

Outline

- 深度学习框架图计算理论
- Beam search
- 生成式文本摘要问题补充
- Baseline代码实践

作业

Bye ！