Generating Sequences with Recurrent Neural Networks - Graves, Alex, 2013

Yuning Mao

Based on original paper & slides

Generation and Prediction

- Obvious way to generate a sequence: repeatedly predict what will happen next
- Best to split into smallest chunks possible: more flexible, fewer parameters

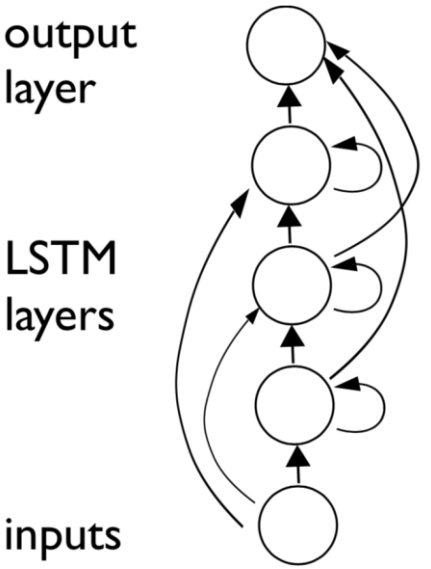
$$\Pr(\mathbf{x}) = \prod_t \Pr(x_t | x_{1:t-1})$$

The Role of Memory

- Need to remember the past to predict the future
- Having a longer memory has several advantages:
  - can store and generate longer range patterns
  - especially ‘disconnected’ patterns like balanced quotes and brackets
  - more robust to ‘mistakes’

Basic Architecture

- Deep recurrent LSTM net with skip connections
- Inputs arrive one at a time, outputs determine predictive distribution over next input
- Train by minimizing log-loss
- Generate by sampling from output distribution and feeding into input



$$\sum_{t=1}^{T} -\log \Pr(x_t | x_{1:t-1})$$

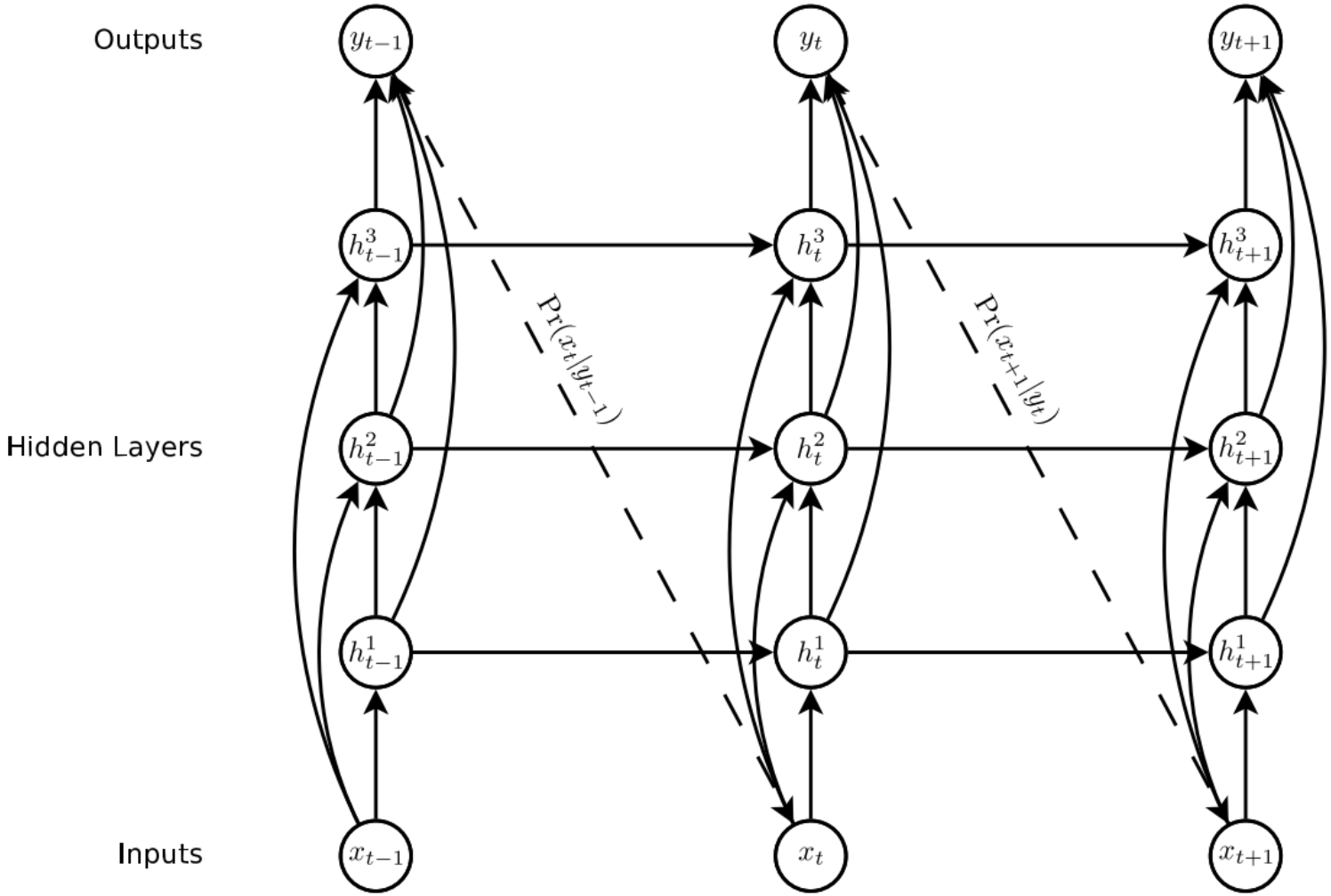
Text Generation

- Task: generate text sequences one character at a time
- Data: raw wikipedia from Hutter challenge (100 MB)
- 205 **one-hot** inputs (characters), 205 way softmax output layer
- Split into length 100 sequences, no resets in between

$$\Pr(x_{t+1} = k|y_t) = y_t^k = \frac{\exp\left(\hat{y}_t^k\right)}{\sum_{k'=1}^{K} \exp\left(\hat{y}_t^{k'}\right)}$$

$$\mathcal{L}(\mathbf{x}) = -\sum_{t=1}^{T} \log \Pr(x_{t+1}|y_t)$$

Network Architecture



$$h_t^1 = \mathcal{H}\left(W_{ih^1}x_t + W_{h^1h^1}h_{t-1}^1 + b_h^1\right)$$

$$h_t^n = \mathcal{H}\left(W_{ih^n}x_t + W_{h^{n-1}h^n}h_t^{n-1} + W_{h^nh^n}h_{t-1}^n + b_h^n\right)$$

$$\hat{y}_t = b_y + \sum_{n=1}^{N} W_{h^n y}h_t^n$$

$$y_t = \mathcal{Y}(\hat{y}_t)$$

Compression Results

Method	Bits per Character
bzip2	2.32
M-RNN[1]	1.6 (text only)
deep LSTM	**1.42** (1.33 validation)
PAQ-8[2]	1.28

1) I. Sutskever et. al. "*Generating Text with Recurrent Neural Networks*" ICML, 2011
2) M. Mahoney, *"Adaptive Weighing of Context Models for Lossless Data Compression"*, Florida Tech. CS-2005-16, 2005

Real Wiki data

'''Albedo''' is the measure of [[reflectivity]] of a surface or body. It is the ratio of [[electromagnetic radiation]] (EM radiation) reflected to the amount incident upon it. The fraction, usually expressed as a percentage from 0% to 100%, is an important concept in [[climatology]] and [[astronomy]]. This ratio depends on the [[frequency]] of the radiation considered: unqualified, it refers to an average across the spectrum of [[visible light]]. It also depends on the [[angle of incidence]] of the radiation: unqualified, normal incidence. Fresh snow albedos are high: up to 90%. The ocean surface has a low albedo. The average albedo of [[Earth]] is about 30% whereas the albedo of the [[Moon]] is about 7%. In astronomy, the albedo of satellites and asteroids can be used to infer surface composition, most notably ice content. [[Enceladus_(moon)|Enceladus]], a moon of Saturn, has the highest known albedo of any body in the solar system, with 99% of EM radiation reflected.

Human activities have changed the albedo (via forest clearance and farming, for example) of various areas around the globe. However, quantification of this effect is difficult on the global scale: it is not clear whether the changes have tended to increase or decrease [[global warming]].

The "classical" example of albedo effect is the snow-temperature feedback. If a snow covered area warms and the snow melts, the albedo decreases, more sunlight is absorbed, and the temperature tends to increase. The converse is tr

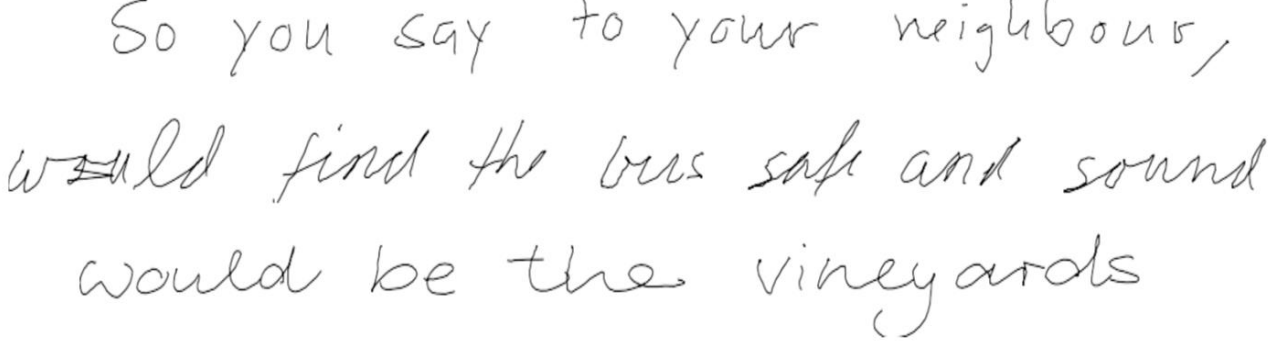
Internet traditions sprang east with [[Southern neighborhood systems]] are improved with [[Moatbreaker]]s, bold hot missiles, its labor systems. [[KCD]] numbered former ISBN/MAS/speaker attacks "M3 5", which are saved as the ballistic misely known and most functional factories. Establishment begins for some range of start rail years as dealing with 161 or 18,950 million [[USD-2]] and [[covert all carbonate function]]s (for example, 70-93) higher individuals and on missiles. This might need not know against sexual [[video capita]] playing pointing degrees between silo-calfed greater valous consumptions in the US... header can be seen in [[collectivist]].

== See also ==

Generated Wiki data

Handwriting Generation

- Task: generate pen trajectories by predicting one (x,y) point at a time
- Data: IAM online handwriting, 10K training sequences, many writers, unconstrained style, captured from whiteboard
- How to predict real-valued coordinates???



Recurrent Mixture Density Networks

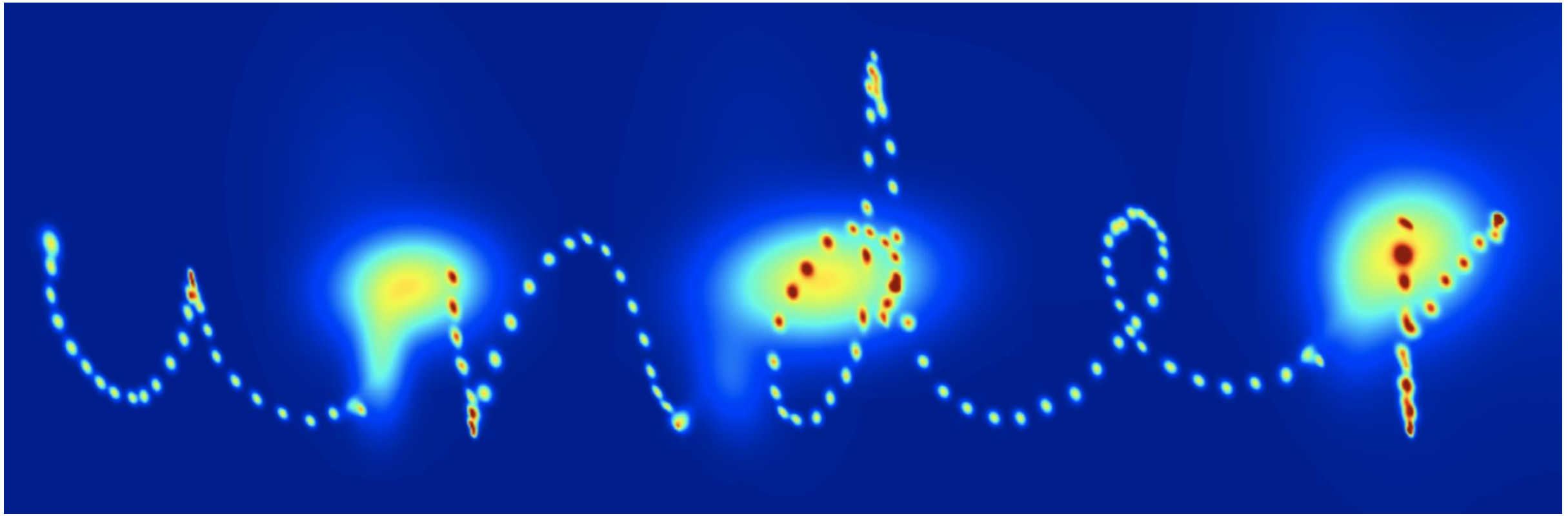
- Suitably squashed output units parameterize a mixture distribution (usually Gaussian)
- Not just fitting Gaussians to data: every output distribution conditioned on all inputs so far
- For prediction, number of components is number of *choices* for what comes next

$$\Pr(o_t) = \sum_i w_i(x_{1:t}) \mathcal{N}\left(o_t | \sigma_i(x_{1:t}), \Sigma_i(x_{1:t})\right)$$

Network Details

- 3 inputs: Δx, Δy, pen up/down
- 121 output units
  - 20 two dimensional Gaussians for x,y = 40 means (linear) + 40 std. devs (exp) + 20 correlations (tanh) + 20 weights (softmax)
  - 1 sigmoid for up/down

$$\Pr(x_{t+1}|y_t) = \sum_{j=1}^{M} \pi_t^j \, \mathcal{N}(x_{t+1}|\mu_t^j, \sigma_t^j, \rho_t^j) \begin{cases} e_t & \text{if } (x_{t+1})_3 = 1 \\ 1 - e_t & \text{otherwise} \end{cases}$$



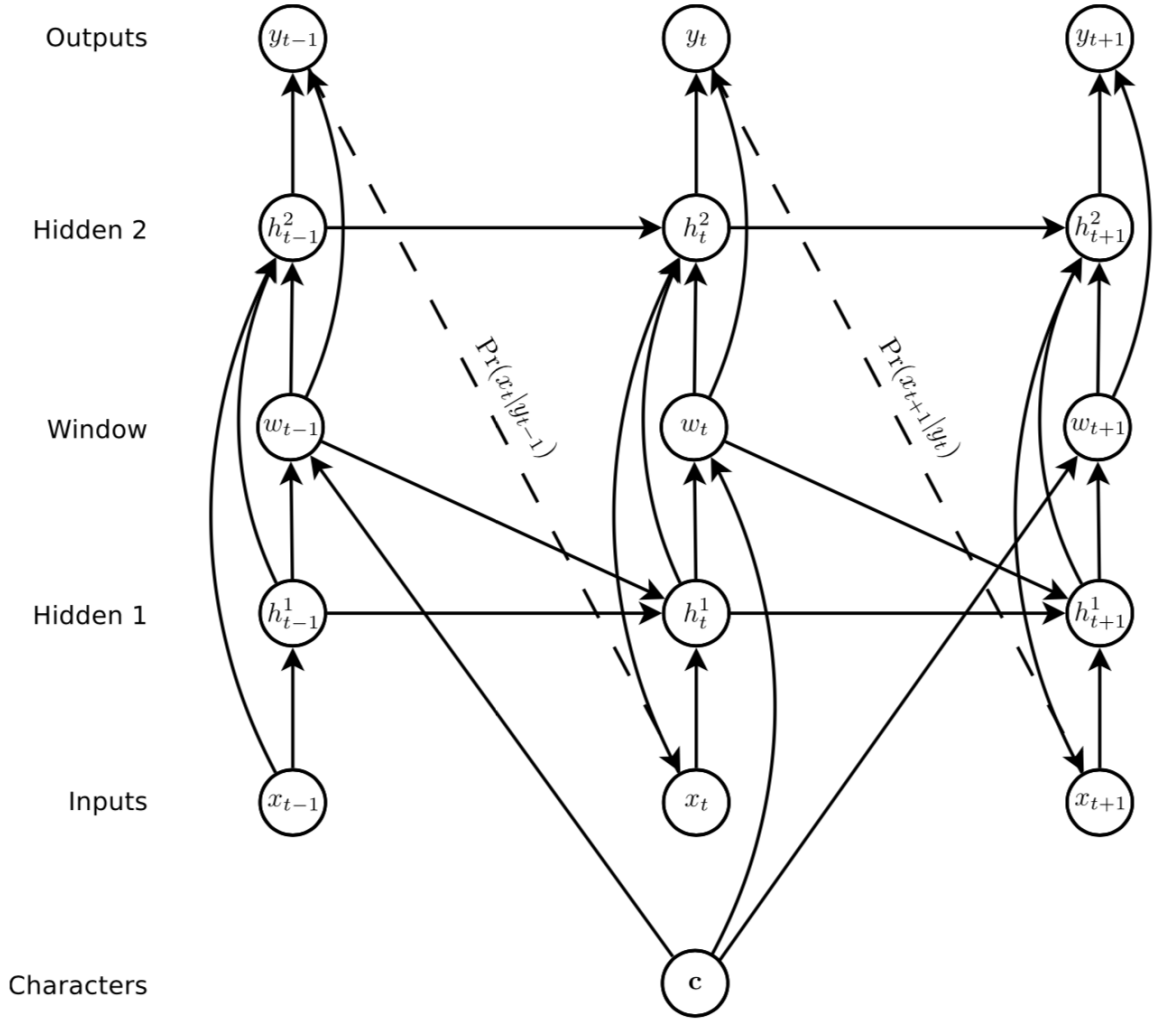
Output Density

Handwriting Synthesis

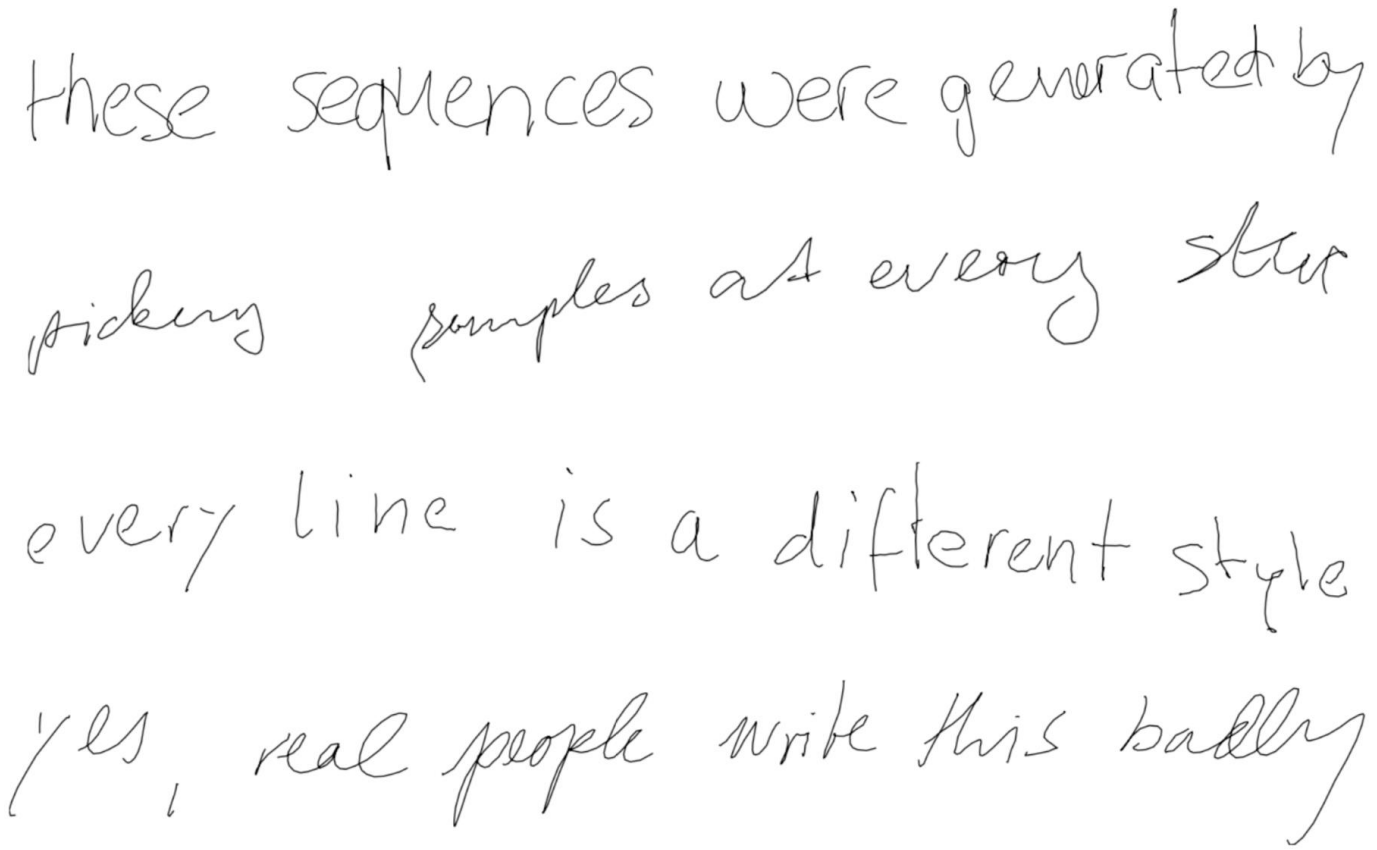
- Want to tell the network *what* to write without losing the distribution over *how* it writes
- Can do this by conditioning the predictions on a text sequence

- Problem: alignment between text and writing unknown
- Solution: before each prediction, let the network decide *where* it is in the text sequence

Network Architecture



Unbiased Sampling



Biased Sampling

when the samples are biased towards more probable sequences they get easier to read but less interesting to look at.