### RNN With Example

Shima Foolad

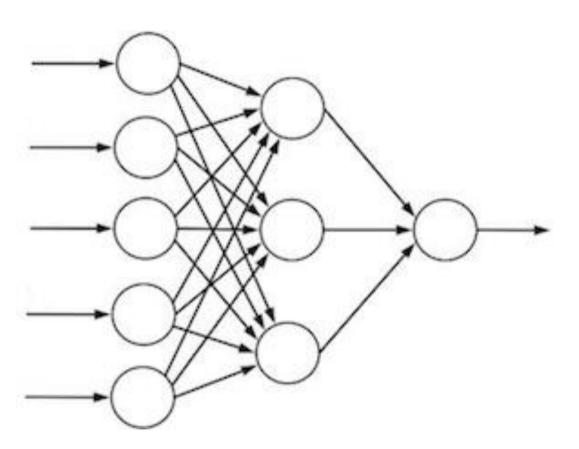
29 Oct. 2017

### RECURRENT NEURAL NETWORK (RNN)

Rohan & Lenny #3: Recurrent Neural Networks & LSTMs Fei-Fei Li & Andrej Karpathy & Justin Johnson: C321n, lecture 10

### ARTIFICIAL NEURAL NETWORK (ANN)

Each neuron stores a single scalar value. Thus, each layer can be considered a vector.



### RECURRENT NEURAL NETWORK (RNN)

each neuron (inputs, hidden(s), and outputs) contains a **vector** of information. The term "cell" is also used, and is interchangeable with neuron.



### RNN (ONE TO ONE)

This is in fact a type of recurrent neural network—a **one to one** recurrent net, because it maps one input to one output. A one to one recurrent net is equivalent to an artificial neural net.



### **EXAMPLE FOR RNN**

Character-level language model example

2017, STANFORD UNIVERSITY, C321N, LECTURE 10 FEI-FEI LI & ANDREJ KARPATHY & JUSTIN JOHNSON

#### min-char-rnn.py gist: 112 lines of Python

```
Minimal character level Vanilla RNN model, Written by Andrej Karpathy (@karpathy)
    838 License
    import numpy as op-
 T Widnes 170
s data = open('input.txt', 'r').read() = smoods be simple plain text file.
y chars - list(set(data))
data_size, vocab_size = len(data), len(chars)
print 'data has Nd characters, Nd unique,' N (data size, vocab size)
char_to_ix = { ch:1 for i,ch in enumerate(chars) }
in_to_char = { i:ch for i,ch in emumerate(chars) }
in w byperparameters
in hidden_size = 100 # size of hidden layer of neurons
If seq_length = 25 * number of steps to unroll the SNN for
learning_rate = ie-1
21 Web = hp.random.randofhiddom_size, vocab_size)*0.81 * imput to hiddom
up Whh = np.random.rando[hidden_size, hidden_size]*6.81 % bidden to hidden
why = np.random.randn(vocab size, hidden size)'6.81 # hidden to nutout
ja bh = np.zeros((hidden_size, 1)) * hidden hims
ju by = np.zeros{(yocah_wize, 1)) * notint hims
    def lossFun(inputs, targets, hprev):
      imputs, targets are buth list of integers.
      horev is Hx1 array of initial hidden state
      returns the loss, gradients on model parameters, and last hidden state
      loss = 0
      # forward pass
      for I in xrange(len(imputs)):
        xs[t] = np.zeros([vocab_size,i]) = econde in 1-of-k representation
        xs[t][inputs[t]] = 1
        hs[t] = np.tanh(np.dot(wsh, xs[t]) + np.dot(wsh, hs[t-1]) + bh) + hidden state
        yx[t] = np.dot(why, hs[t]) + hy + annormalized log probabilities for each chara-
        ps[t] = np.exp[ys[t]] / np.sum(np.exp(ys[t])) = probabilities for rest chara-
        loss += -np.log(ps[t][targets[t],0]) # softmax (cross-entropy loss)
     # Backward pass: compute gradients going backwards
      doon, dwhh, dwhy = pp.zeros_like(Wdh), np.zeros_like(Whh), np.zeros_like(Why)
      dbb, dby = np.zeros_like(bb), np.zeros_like(by)
      dhnext = np.zeros like(hs[0])
      for t in reversed(krange(len(inputs))):
        dy = np.copy(ps[t])
        dy[targets[t]] -= 1 % backgrop into y
        dWhy += np.dot(dy, hs[t]_T)
       dh = np.dot(Mhy.T, dy) + dhnext + backprop into h
       draw = (1 - hs[t] * hs[t]) * dh # backprop through Early nonlinearity
        dwxh == np.dot(dhraw, xs[t].t)
        dWith += op.dot(dbraw, hs[t-1].T)
       dinext - up.dot(Whh.T, dhraw)
      for dparam in [dwxh, dwhh, dwhy, dbh, dby]:
       np.clip(dparam, -s, s, out=dparam) = clip to mitigate employing gradients
      return loss, duch, duch, duby, duby, duy, hs[lan(inputs)-1]
```

```
an def sample(h, seed ix, n):
       sample a sequence of integers from the model.
       h is memory state, seed ix is seed letter for first time step
       x = np.zeros((vocab_size, 1))
       x[seed ix] = 1
       ixes - []
       for t in xrange(n):
         h = np.tanh(np.dot(wxh, x) + np.dot(whh, h) + bh)
         y = np.dot(Why, h) + hy
         p = np.exp(y) / np.sum(np.exp(y))
         ix = np.random.choice(range(vecab_size), p=p.ravel())
         s = np.zeros((vocsb_size, 1))
         x[ix] = 1
         ixes.append(ix)
     mixin, mikh, mikhy = np.zeros like(wikh), np.zeros like(wikh), np.zeros like(why)
     with, why = np.zeros like(bh), np.zeros like(by) = nestry variables for Adapted
 smooth_loss = -np.log(1.8/voceb_size)*seq_length = loss at iteration 0
      * propers imputs (so're sameping from left to right in stops seg_leagth long)
       if p-seq_length+1 >= len(date) or n == 0;
        hprey = np.zeros((hidden size,1)) # reset HHH nemary
        p = 0 = go from start of data
       inputs = [char_to_ix[ch] for ch in data[p:p+seq_length]]
       targets = [char_to_ix[ch] for ch in data[p+1:p+seq_length-1]]
       # sample from the model now and then
      if n s inn == 0:
         sample_ix = sample(hprev, inputs[0], 268)
         txt = ''.join(ix to char[ix] for ix in sample ix)
         print '----\n % \n----' % (Ixt, )
       * forward seq length characters through the net and fetch gradient
       loss, dash, dahh, dahy, dbh, dby, hprev = lossFun(inputs, targets, hprev)
       smooth loss = smooth loss * 0.900 + loss * 0.801
       if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) * print progress
       * perform parameter update with Adagrad
       for param, dparam, now in hip([wxh, whh, why, bh, by],
                                    [dwxh, dwhh, dwhy, dhh, dhy],
1000
                                    [maxh, mahh, mahy, mbh, mby]):
         mes += sparas * sparas
         param += learning_rate * dparam / np.sqrt(mem + ie-8) # adegrad update
111 p += seq_length # abse data pointer
in a += 1 = iteration counter
```

(https://gist.github. com/karpathy/d4dee566867f8291f086)

#### AN EXAMPLE

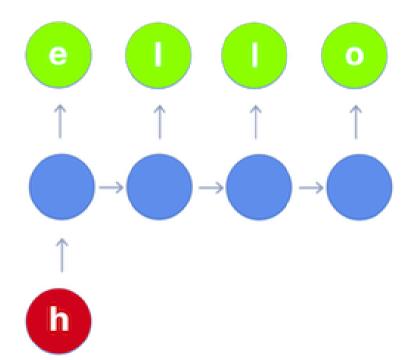
# Character-level language model example

Vocabulary: [h,e,l,o]

Example training sequence: "hello"

"h" = 
$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$
; "e" =  $\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ ; "l" =  $\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$ ; "o" =  $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$ 

This is what we'd expect with a trained RNN:



#### min-char-rnn.py gist

```
for sell rise with these I Would's Mill male! Written by habing Reputly Officepaths
www.doense
coppert respy as he
does a report tracturer; a "financial a woode on a testing hair cost file.
dona him, exceptaint + invidual, instructed
government and the projections, he employed he protection, worst comp
char. no. ic = 1 chil for L ch in commerceleners) !
   to boundary of committee another remarked afficients of
Single size * De o mine of hidden layer of regross
sw_brager - Dr a carbon of binning property the 200 for
 Learning race * 18 1
I book parameters
with a new roader, remark toleranders, wood, size $70.00 at larger to title on
WW TIES, ORGANIZATION CONTRACTOR 
www.e. ac render, reports seem after. Atteden, advertig on a minder on compart
by a salphray("Vecub adpo. '19) a common stars
dof less-un/brown, surgers: horse):
      impacts, cargede are been list of prospers.
       returns the Look, gradients or model parameters, and last misser state
       se. W. Se. PE " Co. Cl. (D. C)
     tof-cl = ex.opeltered
      Jack - a
        fac I in ameger feet reported.
          safet a se sersa (lasera ateria) i a carrie te a ef-a concentrarias
           ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh histor, state)
                                                                                                                                                                                                                                                 data is 'hello'
            will be madel (May, but if it has a common of
            pa[t] = np.eop(ya[t]) + 2.eo.aum(np.eop(ya[t])) + 0.complicates for race observed.
           tex-to-up-loggic/Iffterpris/III and the presentation from
        A DECEMBER DESC. COMBUTE DISCUSTA DOOR DECEMBER.
        each, with, other and arms, I hardenly, optiones, Triciples of agreence, Tabelle
       con, dby " rg.deroe itsetbe; np.meroe itsetby)
       chorac - wa were likelikabila
        for the development and tenchopathy (se
                                                                                                                                                                                                       Vocabulary is
           do + more gray(molth)
           deliamentally and in some or with a
           detry on the decisy, builting
                                                                                                                                                                           'h','e','l','o']
          di + np.doctelly.i. dy9 + dispot + batistrop into n
            thow sittle held foliable for the house trans-
           dry - rym
            charle on the charles are and (1.7)
            duals on application, strict of all
           discut a special older, there;
                                                                                                                                                                                         data size is 5
      for spares on [seen, seen, duty, don, doy];
           op offspelannen, 48, %, destatueren a ettip te mittanen en-
       return took, dean, dam, early, early day, outself security by
   out appliedly seed by the
                                                                                                                                                                                          Vocabulary size is 4
     combined to the property of the property from the shoot
      It is every date, work to be and have for first the every
    > 7 Nel Demonstraction delice. (Er)
Sincrellar) = 1.
face. a []
                                                                                                                                            index of each character:
        is transaction content, of a spinor (Mar, 6) and

    Final contratts, (v) = by
    Expression of the contract tracky)
    A set a contract tracky (contract of the part of the contract of the contr
                                                                                                                                                 {'l':2, 'o':3, 'e':1, 'h':0}
        e ne neus (entré sire 1);
epuil :
         SAME RESIDENCE SAME
    note such such to beyon distances, so been ass
 MET, MON F TO JETCH JERNEY) ME, DOTEL JESTING /
                                                                                                                                    character of each index :
 second_loss a logicogra acressa_consistent_loss; longità salita mass
    or press sectors of sectors of a mile.
                                                                                                                                    {0:'h', 1:'e', 2:'l', 3:'o'}
       have a fator, or, both first or to the frequency from
      respect from to supply for ships delighed prom-
        complete to complete process to many (1), 1970.

Oil - 1 years to the last the action regulator).
        grade ..... who will be to the a
    here, there will place the size terms of the highest compute the contradiction of the contradictions of the c
    day of a past of all prices factors on the past, and a second past of a past of contents.
      for person distance, over its significant bids, was to, og-
                                                                      THERE MOR. MAY, 516 MY HE
        person of the original of the contract of the contract of the contract of the
     piet and Ample 1 room code promise
```

#### Data I/O

```
11 11 11
     Minimal character-level Vanilla RNN model. Written by Andrej Karpathy (@karpathy)
     BSD License
     11 11 11
     import numpy as np
     # data I/O
     data = open('input.txt', 'r').read() # should be simple plain text file
     chars = list(set(data))
     data_size, vocab_size = len(data), len(chars)
     print 'data has %d characters, %d unique.' % (data_size, vocab_size)
11
     char_to_ix = { ch:i for i,ch in enumerate(chars) }
     ix_to_char = { i:ch for i,ch in enumerate(chars) }
```

#### min-char-rnn.py gist Hidden size is 3 Number of timesteps is 4. it means each character is dependent on four character before it. (chunks of 4 characters in a time) WW T NO PRODUCT PRODUCTION OF THE PROPERTY OF THE PROPERTY OF Learning rate is 0.1 www was reader, rescalescentistics, attitude take? O on which is note. 19 wxh is Weight matrix 3\*4 between 20 input layer and output layer 22 whh is Weight matrix 3\*3 between two hidden layers why is Weight matrix 4\*3 between hidden layer and output layer bh is hidden bias by is output bias MET, MON PITA JETOS JERNESNI MELDOTEL LIGHTANI V ROBERT NAVIALISE FOR HOSETAN proper loss a springer access to proper loss of the control of the lapter - Jakes, or, to July fire on the configuration, from \$1. respect from to origin; for all to delegatives in generally simple to the property to accept, 1970, tel. " year to the first just the second set." Now, Chin, Addy, May, Mr., Ang. State. Temple (Injury). Images, Images ments district over its significant tale, was trying! 1860, 860, 866, 03, 87). 1860, 860, 866, 867.

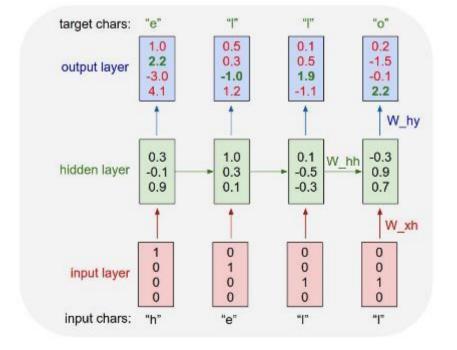
page of Decelopment April 2 (page (new color)) a migration in

#### **Initializations**

```
# hyperparameters
hidden_size = 3  # size of hidden layer of neurons
seq_length = 4  # number of steps to unroll the RNN for
learning_rate = 1e-1

# model parameters
Wxh = np.random.randn(hidden_size, vocab_size)*0.01  # input to hidden
Whh = np.random.randn(hidden_size, hidden_size)*0.01  # hidden to hidden
Why = np.random.randn(vocab_size, hidden_size)*0.01  # hidden to output
bh = np.zeros((hidden_size, 1))  # hidden bias
by = np.zeros((vocab_size, 1))  # output bias
```

#### recall:



#### min-char-rnn.py gist Hidden size is 3

Number of timesteps or sequence length is 4. it means each character is dependent on four character before it. (chunks of 4 characters in a time)

```
Learning rate is 0.1
                                                                                                                     wxh is Weight matrix 3*4 between
                                                                                                                  input layer and output layer
                                                                                                                  whh is Weight matrix 3*3 between
                                                                                                                     two hidden layers
                                                                                                                  why is Weight matrix 4*3 between
                                                                                                                     hidden layer and output layer
                                                                                                                  bh is hidden bias
                                                                                                                  by is output bias
MET, NOT FIRE SETON LIFERENT OR DETER LIFETED FOR THE SET OF THE SETON OF THE SE
        If proof tempters or tempters of a mile.

Somey * Major some respectively. 221, 214 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 1885 - 
        hapter (star outs) at the in-territories trapect request (pare to sept), for all in-deligh-layers trapect().
                complete complete an interest, and the first transfer of the state of 
     News, class, class, class, class, class, terms, terms relations, images, terms a model least a seasoft least a second class and class a second class a second class a second class and class a second cla
                                                                                                                                                                                  1866 858 859 08 87).
1868 858 859 68 8711:
```

#### **Initializations**

```
# hyperparameters
```

hidden\_size = 3 # size of hidden layer of neurons
seq\_length = 4 # number of steps to unroll the RNN for
learning\_rate = 1e-1

#### # model parameters

19

20

22

Wxh = np.random.randn(hidden\_size, vocab\_size)\*0.01 # input to hidden
Whh = np.random.randn(hidden\_size, hidden\_size)\*0.01 # hidden to hidden
Why = np.random.randn(vocab\_size, hidden\_size)\*0.01 # hidden to output
bh = np.zeros((hidden\_size, 1)) # hidden bias
by = np.zeros((vocab\_size, 1)) # output bias

$$W_{xh} = \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix}$$

$$W_{hh} = \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix}$$

$$W_{hy} = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix}$$

$$b_h = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$b_y = [0 \ 0 \ 0 \ 0]$$

2016, STANFORD UNIVERSITY, C321N, LECTURE 10 FEI-FEI LI & ANDREJ KARPATHY & JUSTIN JOHNSON

#### min-char-rnn.py gist

```
For each class with a level West Har Mid. model , No of the by Judicy Regardly (Blangadly)
 www.doense
DIRECT PURPS AS FA
 14140406-1412
  does a report tasks, seed, or the court is woodle by a testing have now falle.
   Chart - Theretake Backet t
 done stor, exem stor a leadannt, leatehoral
 government and the properties, he entire. In printage on work your
   char to ic = 1 chil for Lob in commerce cherry !
   to be about 1 ( ) and he is of the removal above of the
 Stoder spe Tow - size of Stoden lawer of courses
  Societies - 19 property of binary in print the 24 for
   Learning race * 18 1
 I book parenters.
 "wid" + no roader, range (Nicolan, State, Woods, State (N. 00 or larger to Milden
 WW TIES, ORGANIZATION CONTRACTOR 
 www.e. ac render, reports seem after. Atteden, advertig on a minder on compart
 by a salphray("Vecub adpo. '19) a common stars
  def lease-in/trauta, parages, horovit-
        inputs, targets are beth list of incapers.
          returns the Local gradients on eadel parameters, and last misden state
          se. W. Se. PE " Co. Cl. (D. C)
        tof-cl = ex.opeltered
         Jack - a
          fac t in anesected/mostrati-
             safet a se sersa (lasera ateria) i a carrie te a ef-a concentrarias
               ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh histor, state)
                will be madefully, but if a be a communication or an absorbit on the analysis.
                pa[t] = np.exp[ya]t]t \ge ex.exe(np.exp[ya]t]t) = drobabilities for race of are
             tion to an important flamphical for a community of the form
          e papieras para: compute practicata, 20030 percentas.
           each, side, side a open usual harbord, open out including open out the play.
          con, dby " rg.dwroe like(bb), np merce like(by)
          chorac - wa were likelikabila
           for the few medianage respectively opening to
               do + more gray(molth)
               deliamentally and in some or with a
               detry on the decisy, builting
             di a se derigio, i. del a discor a batterior fatte in
               those sign and the second second second second second second
               dry - rym
               charle on the charles are and (1.7)
               dean or sp. doc|cares. na(t-s).r)
               discut a special older, there;
         for spares on [seek, sent, dety, don, doy]:
               op stigetheres, 45 %, orreduced a city to attach confedent and tiers
         return seek, dean, dam, early, ean, day, begandpackets) 13
   per admitted, seed the first
      combined an income of terrorise from the south
         I however, said, work to be and have the first time one
      > 7 Nel Demonstraction delice. (Er)
Sincrellar) = 1.
face. a []
           is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

           a national (soletation 1);
april 12
            SAME RESIDENCE SAME
   MET, NOW A TO LEGISLES HEADY NELDONES, LICEUSE A RESERVE ANGULAR THE ACRESS
   second_loss = log logic strend_scriptos_logic = loss of relevative while receive
      or oregularites - telligital of e = 0
      tigens (place on to add) for or to design areas, trapical respect to add), for all in design types in generally
          complete to stop before the money, 1990
bit "year to the fact the sector contents"
must be with the " in this, i
      Then, Ching (MA), MA, MR, Ang. Server. Second a Jupine, Language, Language, annuals lates a second lates a 2,000 or Language (LANGUAGE). The a page 1000 of Language (LANGUAGE) and a second lates a seco
          he seems district over its significant bits, was to, og-
                                                                                 TIME BOD, BOD, COL BUILD THE BOD OF THE STATE OF THE STAT
           person in the rating ratio 1 dynamic appropriate a recall in migration to
      p = geq leggs i note cuts pourse.
8 = 5 s incrette increte.
```

#### Main loop

```
n, p = 0, 0
81
     mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
     mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
     smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
     while True:
       # prepare inputs (we're sweeping from left to right in steps seg length long)
       if p+seq length+1 >= len(data) or n == 0:
87
88
         hprev = np.zeros((hidden size,1)) # reset RNN memory
         p = 0 # go from start of data
89
       inputs = [char_to_ix[ch] for ch in data[p:p+seq_length]]
91
       targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
       # sample from the model now and then
       if n % 100 == 0:
94
         sample_ix = sample(hprev, inputs[0], 200)
         txt = ''.join(ix_to_char[ix] for ix in sample_ix)
97
         print '----\n %s \n----' % (txt. )
99
       # forward seq_length characters through the net and fetch gradient
       loss, dWxh, dWhh, dWhy, dbh, dby, hprev = lossFun(inputs, targets, hprev)
100
       smooth loss = smooth loss * 0.999 + loss * 0.001
101
       if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) # print progress
102
103
       # perform parameter update with Adagrad
104
       for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
105
106
                                      [dWxh, dWhh, dWhy, dbh, dby],
107
                                      [mWxh, mWhh, mWhy, mbh, mby]):
         mem += dparam * dparam
108
         param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
109
110
       p += seq_length # move data pointer
111
112
       n += 1 # iteration counter
```

#### min-char-rnn.py gist

```
Normal viscousing Jesus Months MA majet, Northworld Subseq Regardly (Strengtsby
            www.doense
           DIRECT PURPS AS FA
             does a report tasks, seed, or the court is woodle by a testing have now falle.
            dona stor, excessator e limitadat, lesteneral
            government and the properties, he entire. In printage on work your
              char to ic = 1 chil for Lob in commerce cherry !
              to be about 1 ( ) and he is of the removal above of the
            Sictor size * De o size of Middle layer of regrand
             SN_PRICE - IN Electric of Street or social the 200 for
              Learning race * 18 1
            I book prosesses
             with a new roader, remark toleranders, wood, size $70.00 at larger to title on
            WW TIES, ORGANIZATION CONTRACTOR 
            www.wo.rowdor.rowonlescop.aige. httdcs_nigetY0.oi = nidden in compet.
            for a sacrement (week, adject 101) a common
             def lease-in/trauta, parages, horovit-
                                                                                                                           initial hidden vector 3*1 to zero
                   impace, cargete are been list of prospers
consider indices of data with sequence length
as inputs
                    A DECEMBER DESC. COMBUTE DISCUSTA DOOR DECEMBER.
                    each, with, who a open usual harmon, opening harpoon, opening the plays.
                    con, dby " rg.deroc like(bb), np.deroc like(by)
                    chorac - wa were likelikabila
                    for the few medianage respectively opening to
                       do + more gray(molth)
                       detrangenetation of the secretary arms of
                       detry on the decisy, builting
                      di a se derigio, i. del a discor a batterior fatte in
                       those sign and the second second second second second second
                       dry - rym
                       charle on the charles are and (1.7)
                       dean or sp. doc|cares. na(t-s).r)
                       discut - epications T, these;
                   for spares on [seek, sent, dety, don, doy]:
                       op stigetheres, 45 %, orreduced a city to attach confedent and tiers
                   return seek, dean, dam, days, ean, day, highertassatis, sig-
              per admitted, seed the first
                  combined to the property of the property from the shoot
                    I have you and wish to be and being the first time of
                  or hat partners worst state. The
                   tors a III
                    is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

                    a national (soletation 1);
april 12
                     SAME RESIDENCE SAME
                MET, MON P TO JETTO JERNETH MELDEREL LIGHTEN / MONEY ANGLESS FOR ASSETS
              server_loss a vag lagra alterna_straines_tomph a loss or interior shifts from
                 or oregularites - telligital of e = 0
                  respect from to supply for their delighed grown in growith
                    complete to stop before the money, 1990
bit "year to the fact the sector contents"
must be with the " in this, i
                 Then, Ching (Mill), Milly, Mill, Milly, Norway, North (Ching) in , Imprily, Immediate the action of the Ching (North Lance (North Lance)) and the ching (North Lance) and 
                    he seems district over its significant bits, was to, og-
                                                                             THERE MADE, MAY, EDG. MY HE
                    person in the original Colored Colored (see a sea) is selected as a
```

#### Main loop

```
n, p = 0, 0
                   mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
                   mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
                   smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
                   while True:
                     # prepare inputs (we're sweeping from left to right in steps seg length long)
                     if p+seq length+1 >= len(data) or n == 0:
               87
                       hprev = np.zeros((hidden size,1)) # reset RNN memory
                       p = 0 # go from start of data
               89
                     inputs = [char to ix[ch] for ch in data[p:p+seq length]]
               91
                     targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
                     # sample from the model now and then
                     if n % 100 == 0:
               94
                        sample_ix = sample(hprev, inputs[0], 200)
in our example, data is 'hello' and consider indices of 4 characters
            (sequence length is 4) of data as inputs ('hell') .
                                                                                       hprev)
                         then inputs are [0, 1, 2, 2]
                                                                                      rint progress
                     for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
              105
              106
                                                    [dWxh, dWhh, dWhy, dbh, dby],
              107
                                                    [mWxh, mWhh, mWhy, mbh, mby]):
                       mem += dparam * dparam
              108
                       param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
              109
              110
                     p += seq_length # move data pointer
              111
```

112

n += 1 # iteration counter

he seems district over its significant bits, was to, og-

THE MODE WAY, DOG TO HE !

person in the original Colored Colored (see a sea) is selected as a

```
min-char-rnn.py gist
         Normal viscousing Jesus Months MA majet, Northworld Subseq Regardly (Strengtsby
         www.doense
         DREAT PURPO AS FA
         does a report tract, seri, in the rest is woode as a testing from cost fills.
         dona stor, excessator e limitadat, lesteneral
         good risks as as paracters, as anger? a paracter, weak your
          char to ic = 1 chil for Lob in commerce cherry !
          to be about 1 ( ) and he is of the removal above of the
         Sictor size * De o size of Middle layer of regrand
         SN_PRICE - IN Electric of Street or social the SN for
          Learning race * 18.3
         I book prosesses
         with a new roader, remark toleranders, wood, size $70.00 at larger to title on
         With Tigo (Wildows Proceedings on No. 1920) 1979; T. H. & Partier for Trader
         www.enc.rendor.rendolescen.atec. Atedes.atec? (0.00 = 10.00 to a consti-
         for a sacrement (week, adject 101) a common
          dof lease-in/trauta, parages, horovit
                                                                                       initial hidden vector 3*1 to zero
              impacs, cargede are been list of prospers
consider indices of data with sequence length
as inputs
the next characters of each character inputs are
targets
                                       held I hap It has been as transported and search
                 charle on the charles are and (1.7)
                 duals on application, strict of all
                discut - epications T, these;
            for sparse in [sect. serv. day, day, day]; 
op clipidparen, -6, 5, der-derice) o clip to mittanic confedibil produces.
             return seek, dean, dam, early, ean, day, begandpackets) 13
           out appliedly seed by the
             combined to the property of the property from the shoot
              I have you and wish to be and being the first time of
             S. F. SELENTONTHOUSE SIZE, 37 (
             tors a III
              is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

              a national (soletation 1);
april 12
               SAME RESIDENCE SAME
           MET, MON P TO JETTO JERNETH MELDEREL LIGHTEN / MONEY ANGLESS FOR ASSETS
          proset_loss a .eq.logs.styeos_totales_losses i totales controlled
            or oregularites - telligital of e = 0
             respect to the solution of an ability-disperse in generally
              simple in a spinite man a second (1) (20)

of a "grant of the last two as an electrical and

most a way to the man of the columns and a
            Then, Ching (Mill), Milly, Mill, Milly, Norway, North (Ching) in , Imprily, Immediate the action of the Ching (North Lance (North Lance)) and the ching (North Lance) and
```

81

87

89

91

94

105

106

107

108

109 110

111

112

#### Main loop

```
p = 0, 0
                mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
                mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
                smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
                while True:
                  # prepare inputs (we're sweeping from left to right in steps seg length long)
                  if p+seq_length+1 >= len(data) or n == 0:
                    hprev = np.zeros((hidden size,1)) # reset RNN memory
                    p = 0 # go from start of data
                  inputs = [char to ix[ch] for ch in data[p:p+seq length]]
                  targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
                  # sample from the model now and then
                  if n % 100 == 0:
                    sample_ix = sample(hprev, inputs[0], 200)
in our example, input is 'hell', target is 'ello' and consider its
                                 indices.
                                                                                    hprev)
                       then inputs is [1, 2, 2, 3]
                                                                                   rint progress
                  for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
                                                [dWxh, dWhh, dWhy, dbh, dby],
                                                [mWxh, mWhh, mWhy, mbh, mby]):
                    mem += dparam * dparam
                    param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
                  p += seq_length # move data pointer
                  n += 1 # iteration counter
```

#### min-char-rnn.py gist

```
No sed country-level Worlds MA solel, Written by Johny Republy (Brognitty)
www.doense
DREAT PURPO AS FA
 does a report tract, seri, in the rest is woode as a testing from cost fills.
dona stor, excessator e limitadat, lesteneral
government and the properties, he entire. In printage on work your
  char to ic = 1 chil for Lob in commerce cherry !
  to be about 1 ( ) and he is of the removal above of the
Miches size * Day - size of Miches layer of regrand
 SN_PRICE - IN Electric of Street or social the SN for
  Learning race * 18 1
I book prosesses
 with a new roader, ramen@tologic.com, wood_vizeCV0.00 in tright to hillion.
WW TIES, ORGANIZATION CONTRACTOR 
www.e. ac render, reports seem after. Atteden, advertig on a minder on compart
by a salphray("Vecub adpo. '19) a common stars
 def lease-in/trauta, parages, horovit-
      inputs, targets are beth list of incapers.
        returns the Local gradients on eadel parameters, and last misden state
        se. W. Se. PE " Co. Cl. (D. C)
      tof-cl = ex.opeltered
       Jack - a
        far I be arresected reaction.
            safet a se sersa (lasera ateria) i a carrie te a ef-a concentrarias
             ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh histor, state)
              will be madefully, but if a be a communication or an absorbit on the analysis.
              pa[t] = np.exp[ya]t]t \ge ex.exe(np.exp[ya]t]t) = drobabilities for race of are
             tex-to-up-begindiffleepitidif, bij o' of our procurement to-i
          e papierana basi, sombite bisconsyt 300,00 periesaana
          each, side, side a open usual harbord, open out including open out the play.
        con, dby " rg.dwroe like(bb), np merce like(by)
        chorac was owns. Mississists
          for the few medianage respectively opening to
             do + more gray(molth)
             detrangenetation of the secretary arms of
             detry on the decisy, builting
            di a se derigio, i. del a discor a batterior fatte in
              thow off a held foliable for the beautiful transported and search
             dry - rym
              charle on the charles are and C. T.
              duals on application, strict of all
             discut - epications T, these;
        fit spares in [sect. sent, dety, don. dog:
             op stigetheres, 45 %, orreduced a city to attach confedent and tiers
        return took, dean, dam, days, ean, day, better(basets) 13
            Do a forward pass, backward pass, and get the
              gradients in a Loss function (lossFun)
           SAME, MESSAGES SAV
    MET, MON P TO JETTO JERNETH MELDEREL LIGHTEN / MONEY ANGLESS FOR ASSETS
  second_loss = log logic strend_scriptos_logic = loss of relevative while receive
     or oregularites - telligital of e = 0
     tigens (place on to add) for or to design areas, trapical respect to add), for all in design types in generally
        complete to stop before the money, 1990
bit "year to the fact the sector contents"
must be with the " in this, i
     Then, Ching (MA), MA, MR, Ang. Server. Second a Jupine, Language, Language, annuals lates a second lates a 2,000 or Language (LANGUAGE). The a page 1000 of Language (LANGUAGE) and a second lates a seco
        he seems district over its significant bits, was to, og-
                                                                             TIME BOD, BOD, COL BUILD THE BOD OF THE STATE OF THE STAT
          person in the rating ratio 1 dynamic appropriate a recall in migration to
```

#### Main loop

81

87

88

89

91

94

97

103

104

105

106

107

108

109

111

112

```
n, p = 0, 0
mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
while True:
  # prepare inputs (we're sweeping from left to right in steps seg length long)
  if p+seq length+1 >= len(data) or n == 0:
    hprev = np.zeros((hidden size,1)) # reset RNN memory
    p = 0 # go from start of data
  inputs = [char to ix[ch] for ch in data[p:p+seq length]]
  targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
  # sample from the model now and then
  if n % 100 == 0:
    sample_ix = sample(hprev, inputs[0], 200)
    txt = ''.join(ix_to_char[ix] for ix in sample_ix)
    print '----\n %s \n----' % (txt. )
  # forward seq_length characters through the net and fetch gradient
  loss, dWxh, dWhh, dWhy, dbh, dby, hprev = lossFun inputs, targets, hprev)
  smooth loss = smooth loss * 0.999 + loss * 9.00
  if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) # print progress
  # perform parameter update with Adagrad
  for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
                                 [dWxh, dWhh, dWhy, dbh, dby],
                                 [mWxh, mWhh, mWhy, mbh, mby]):
    mem += dparam * dparam
    param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
  p += seq_length # move data pointer
  n += 1 # iteration counter
```

#### min-char-rnn.py gist

```
We not viscously dead benefit with the Mill matel. Mr. Clark to believe because discounted
www.doense
DISCOUNT FAMILY AS NO
 does - specifications; or himself - woods as about glob over file
dona stor, excessator e limitadat, lesteneral
group rists as an electrical, an entire? a particular, worst pro-
  char to ic = 1 chil for Lob in commerce cherry !
  to be about 1 ( ) and he is of the removal above of the
Stoder size * Day o play of Stoder lawer of Approva
 SN_PRICE - IN Electric of Street or social the SN for
  Learning race * 18.3
I book prosesses
"wid" + no roader, range (Nicolan, State, Woods, State (N. 00 or larger to Milden
With Tigo (Wildows Proceedings on No. 1920) 1979; T. H. & Partier for Trader
www.enc.rendor.rendolescen.atec. Atedes.atec? (0.00 = 10.00 to a consti-
for a salebreal (versit adject to), a same about
 def less-un/browns, pargets, horseld
     impacts, cargede are been list of incopers.
      returns the Local gradients on model parameters, and last misden state
    tof-cl = ex.opeltered
     Jack - a
        safet a se sersa (lasera aler, a)) a carele to a of a consequence of
        ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh historistical)
         pa[t] = ap.exp[ya]t]t > ex.exe(ap.exp[ya]t]t) + probabilization for more observed.
        tex-to-up-higher place that a solution to make the first
      a majorant pass: compare gracients gody; mechanics
state, state, state, only a sp. state, likelikolj, og serve, likelikolj, og serve, likelikolj
      con, dby " rg.dwroe like(bb), np.merce like(by)
      chorac was owns. Mississists
      for the few mediarespectancepatry():
        do + mo.con/(soft)
        Astronomically 271 Programs who is
        den in m. desiry, hills. it
        di + no doctaby. . . dy9 + dincot + bapairos data n
         thow will a health happy had a hadron transported and real manage
        dry or ryme
        charle on the charles are profit I TI
        0/95 -- 10.000 05784. 11(1-1).71
        dieset - epicatout T, these;
     for spares as [seen, seen, dety, don: styl):
        op stigetheren. -6, 5, our-decree a city to account contribut gradients
      return took, dean, darm, darry, ear, day, bullestancetts til
    combined to the property of the property from the shoot
    S. F. SELENTONTHOUSE SIZE, 37 (
     tors a III
      is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

      a national (soletation 1);
april 12
       SAME RESIDENCE SAME
  MET, NOW A THE LEGISLE CONTROL AND DESCRIPTION OF A PARTY AND ADDRESS TO A ADDRESS.
 security loss a log logic absentuation long to the accommodate many white recent
    If 0-eeq leggs-1 = 10 lister 0 0 = 0
morey = 00.0010010010010.001, 1 1 = 1000 = 00.0007
      tigues - fator, or, to pay the entire to the Broatenay, he gottle
    respect from to supply for ships delighed grown to growing
      complete to complete process of managers, 1970.

Oil - 1 years of the size (and the second contracts).
      grade ..... who will be to the a
    here, there will place the size terms of the highest compute the contradiction of the contradictions of the c
    into a passes at print later on Later with a few assets into a print concrete
     for seven directs, we in significant tile, was try tell
                                                 THE MODE WAY, DOG TO VILL
      person in the original Colored Colored (see 1964) is selected units
```

#### Loss function

Forward pass

- forward pass (compute loss)
- backward pass (compute param gradient)

```
lossFun(inputs, targets, hprev):
       inputs, targets are both list of integers.
29
       hprev is Hx1 array of initial hidden state
31
       returns the loss, gradients on model parameters, and last hidden state
       xs, hs, ys, ps = {}, {}, {}, {}
      hs[-1] = np.copy(hprev)
34
       loss = 0
       # forward pass
       for t in xrange(len(inputs)):
        xs[t] = np.zeros((vocab_size,1)) # encode in 1-of-k representation
         xs[t][inputs[t]] = 1
        hs[t] = np.tanh(np.dot(Wxh, xs[t]) + np.dot(Whh, hs[t-1]) + bh) # hidden state
        ys[t] = np.dot(Why, hs[t]) + by # unnormalized log probabilities for next chars
42
         ps[t] = np.exp(ys[t]) / np.sum(np.exp(ys[t])) # probabilities for next chars
         loss += -np.log(ps[t][targets[t],0]) # softmax (cross-entropy loss)
44
       # backward pass: compute gradients going backwards
       dWxh, dWhh, dWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
       dbh, dby = np.zeros like(bh), np.zeros like(by)
       dhnext = np.zeros_like(hs[0])
       for t in reversed(xrange(len(inputs))):
         dy = np.copy(ps[t])
        dy[targets[t]] -= 1 # backprop into y
         dWhy += np.dot(dy, hs[t].T)
         dby += dy
         dh = np.dot(Why.T, dy) + dhnext # backprop into h
54
         dhraw = (1 - hs[t] * hs[t]) * dh # backprop through tanh nonlinearity
         dbh += dhraw
         dWxh += np.dot(dhraw, xs[t].T)
         dWhh += np.dot(dhraw, hs[t-1].T)
        dhnext = np.dot(Whh.T, dhraw)
       for dparam in [dwxh, dwhh, dwhy, dbh, dby]:
```

np.clip(dparam, -5, 5, out=dparam) # clip to mitigate exploding gradients

return loss, dWxh, dWhh, dWhy, dbh, dby, hs[len(inputs)-1]

```
min-char-rnn.py gist
                                                                                                                                                                                                                                 def lossFun(inputs, targets, hprev):
  www.doense
  Disport number as he
                                                                                                                                                                                                          29
                                                                                                                                                                                                                                         inputs, targets are both list of integers.
  does - specifications; or himself - woods as about glob over file
                                                                                                                                                                                                                                         hprev is Hx1 array of initial hidden state
  Sona high, execution a limitional, leateneral
  good risks as as paracters, as anger? a paracter, weak your
    cher. to ic = { ch:1 for 1 ch in cramerate|chers) }
                                                                                                                                                                                                                                         returns the loss, gradients on model parameters, and last hidden state
                                                                                                                                                                                                          31
    to to star in a constitution of the research of the control of
  Single size + See o size of Sizes Layer of regrees Sections - 19 a contact of three contact of the 24 for
                                                                                                                                                                                                                                          11 11 11
                                                                                                                                                                                                          32
    Learning race = 18 3
                                                                                                                                                                                                          33
                                                                                                                                                                                                                                        xs, hs, vs, ps = {}, {}, {}, {}, {}
   with a new teacher, remain@bloker.pdect_woods, viceCF0, 00, with part to fill like
   WIN THE REPORT OF THE PROPERTY OF THE PROPERTY
   www.wo.norder.nordelescon.atec. atodes.atec? (Con a ninter in country
                                                                                                                                                                                                          34
                                                                                                                                                                                                                                         hs[-1] = np.copy(hprev)
  by a salebrasi (week adec. 10), a commit also
    def lease-indicates, parages, horovita
                                                                                                                                                                                                          35
                                                                                                                                                                                                                                         loss = 0
      impacts, cargede are both list of incapart
                                                                                                                                                                                                          36
                                                                                                                                                                                                                                         # forward pass
      returns the loss, gradients or eadel parameters, and last mister state
                                                                                                      for each timestep
      tof-cl - ex.comilhered
                                                                                                                                                                                                                                         for t in xrange(len(inputs)):
                                                                                                                                                                                                          37
      Jack - a
                                                                                                                                                                                                                                                  xs[t] = np.zeros((vocab_size,1)) # encode in 1-of-k representation
           Convert our input character at this
                                                                                                                                                                                                                                                  xs[t][inputs[t]] = 1
           timestep to a one-hot vector
                                                                                                                                                                                                                                                  hs[t] = np.tanh(np.dot(Wxh, xs[t]) + np.dot(Whh, hs[t-1]) + bh) # hidden state
                                                                                                                                                                                                          40
                                                                                                                                                                                                                                                  ys[t] = np.dot(Why, hs[t]) + by # unnormalized log probabilities for next chars
                                                                                                                                                                                                          41
        detrangenetally for account other
        down in to desiry highly
                                                                                                                                                                                                          42
                                                                                                                                                                                                                                                  ps[t] = np.exp(ys[t]) / np.sum(np.exp(ys[t])) # probabilities for next chars
        di = mp.doctahy... dy9 + dimos + betrareg into n
        thow off - hold f haply to be a believe transported and reality
        dia -- cares
                                                                                                                                                                                                                                                   loss += -np.log(ps[t][targets[t],0]) # softmax (cross-entropy loss)
        charle on the charles are profit T. T.
        0/95 - 10.005 05"W4. 11[1-1].11
        these - reminister, these
      for spares on [seek, seek, daily, dot; doy];
        republication ent. 48, 35, detectorization e chia to mili-
       reterm seek, dean, damn, early, early only, helpertainents; til
       perfect the second of terror to four the south
                                                                                  h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)
                                                                                                                                                                                                                                                                                                                                                                                                                                 Softmax classifier
                                                                                                           =W_{hy}h_t+b_{v}
                                                                                                                                                                                                                                                                                                                                                                                                                                    L_i = -log(p_{target_i})
      or change designation of the factor of the first control of the factor o
       tigates. [also outlied at the configuration, traped]
        probability
        distribution
```

2016, STANFORD UNIVERSITY, C321N, LECTURE 10 FEI-FEI LI & ANDREJ KARPATHY & JUSTIN JOHNSON

(Softmax)

timestep 1: (t=1)

Example training sequence: "hello"

$$W_{hh} = \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix} \quad W_{xh} = \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix}$$
$$b_h = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

hidden layer
$$h_{t}$$

$$W_{xh}$$

$$0.5$$

$$0.4$$

$$W_{xh}$$

$$1$$

$$0$$

$$0$$

$$0$$
input layer
$$x_{t}$$

$$0$$

$$0$$
input chars: "h"

$$h_{1} = tanh \left( \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \right)$$

$$= tanh \left( \begin{bmatrix} -0.5 \\ 0.5 \\ 0.4 \end{bmatrix} \right) = \begin{bmatrix} -0.4 \\ 0.5 \\ 0.4 \end{bmatrix}$$

Example training sequence: "hello"

$$W_{hy} = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix} \qquad b_y = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$$

output layer
$$y_{t}$$

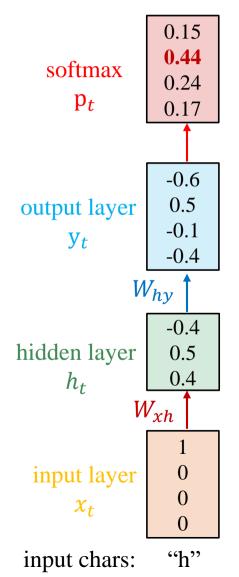
$$v_{t}$$

$$y_t = W_{hy}h_t + b_y$$

$$y_1 = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix} \begin{bmatrix} -0.4 \\ 0.5 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.6 \\ 0.5 \\ -0.1 \\ -0.4 \end{bmatrix}$$

### timestep 1:

(t=1)

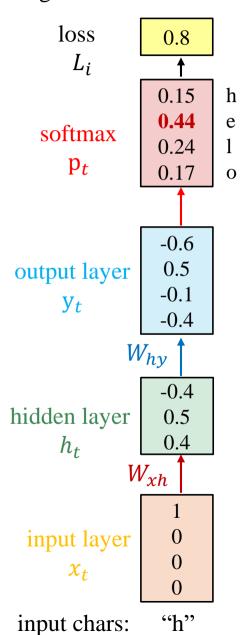


$$p_t = \frac{e^{y_t}}{\sum_j e^{y_j}}$$

target chars:

timestep 1:

(t=1)



Example training sequence: "hello"

$$L_i = -log(p_{target_i})$$

$$L_1 = -log(0.44) = 0.8$$

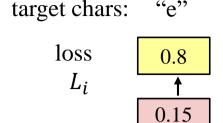
$$L_1 = -log(0.44) = 0.8$$

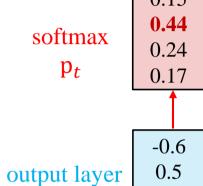
probability distribution of target char (e)

target chars: "e"

timestep 2:

(t=2)





hidden layer 
$$0.5$$
  $0.4$   $0.3$   $0.1$   $0.4$   $0.3$   $0.1$   $0.4$   $0.3$   $0.1$   $0.4$   $0.3$   $0.1$   $0.4$   $0.3$   $0.1$   $0.4$   $0.3$   $0.1$   $0.1$ 

"h"

input chars:

"e"

Example training sequence: "hello"

$$W_{hh} = \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix} \qquad W_{xh} = \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix}$$

$$b_h = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$$

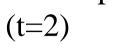
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

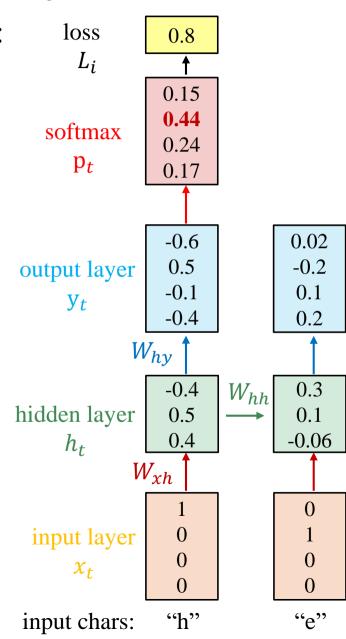
$$\begin{vmatrix} h_2 = tanh \begin{pmatrix} \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix} \begin{bmatrix} -0.4 \\ 0.5 \\ 0.4 \end{bmatrix} + \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.3 \\ 0.1 \\ 0.06 \end{bmatrix}$$

target chars: "e"

timestep 2:





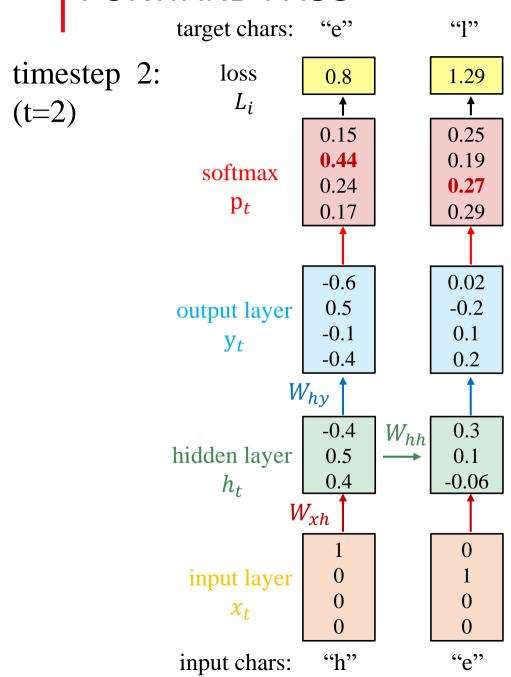
Example training sequence: "hello"

$$W_{hy} = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix} \qquad b_y = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$$

$$y_t = W_{hy}h_t + b_y$$

$$y_2 = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.1 \\ -0.06 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.02 \\ -0.2 \\ 0.1 \\ 0.2 \end{bmatrix}$$

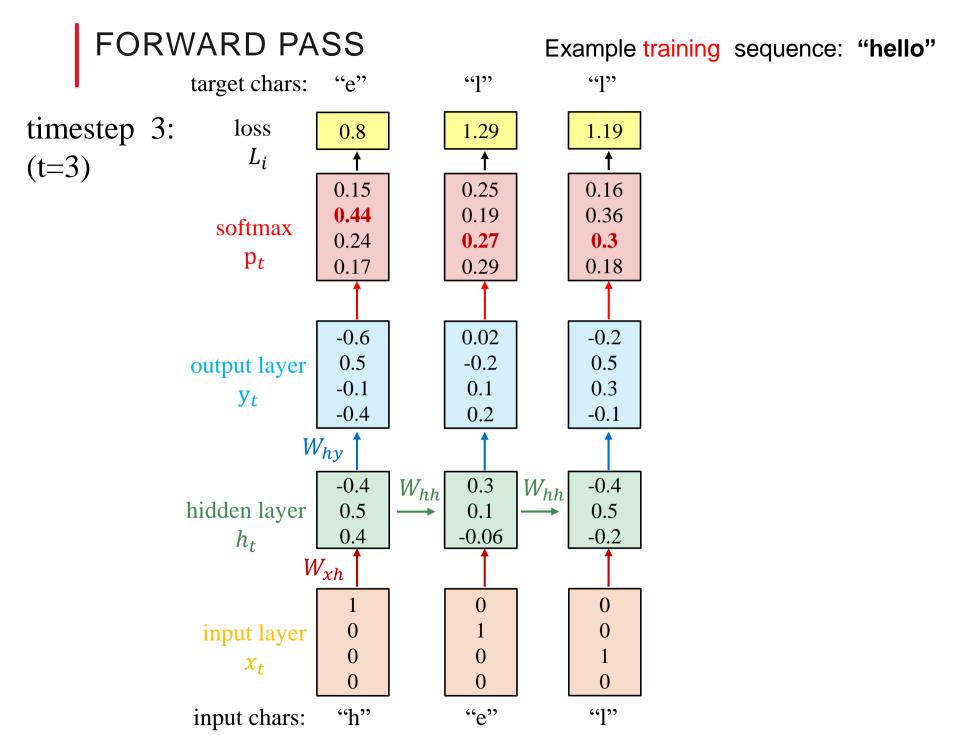
Example training sequence: "hello"

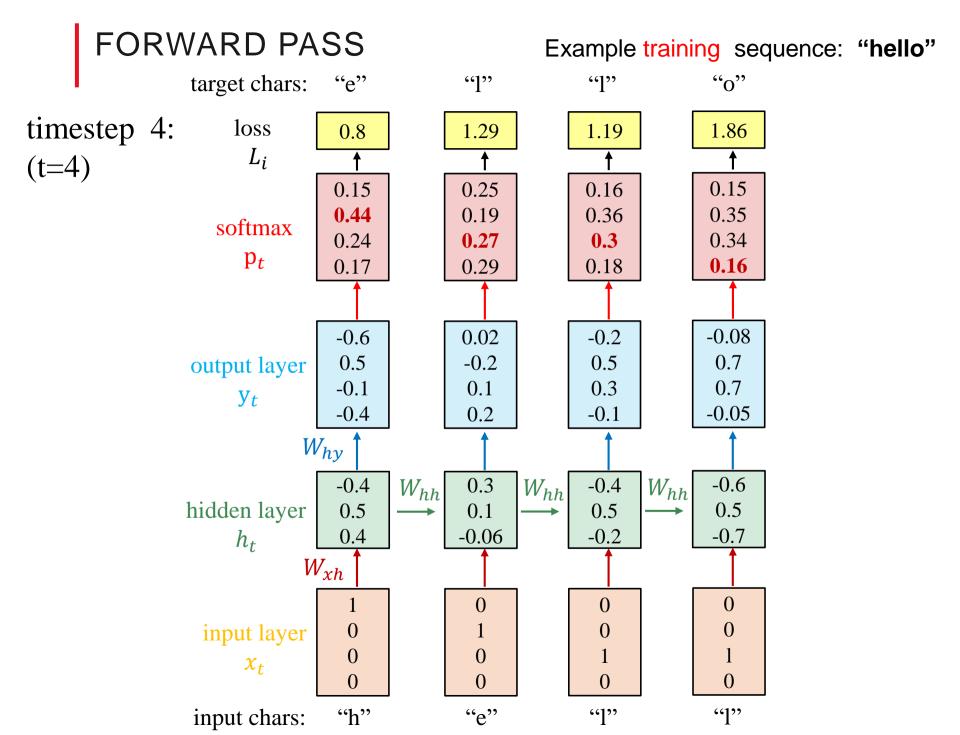


$$L_i = -log(p_{target_i})$$

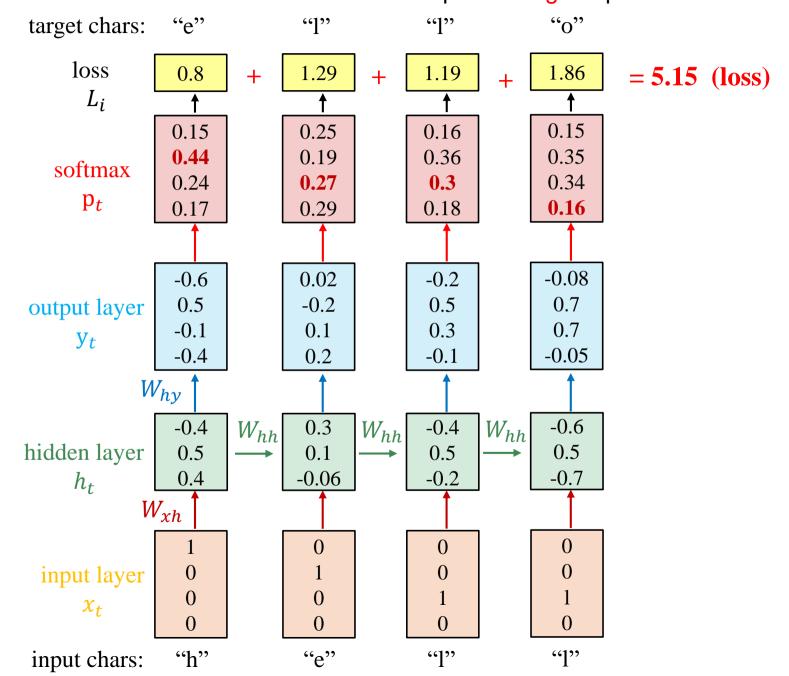
$$L_1 = -log(0.27) = 1.29$$

probability distribution of target char (1)





Example training sequence: "hello"



#### min-char-rnn.py gist

```
Normal viscousing Jesus Months MA majet, Northworld Subseq Regardly (Strengtsby
www.doense
DISCOUNT FAMILY AS NO
 does - specifications; or himself - woods as about glob over file
dona stor, excessator e limitadat, lesteneral
group rists as an electrical, an entire? a particular, worst pro-
  char to ic = 1 chil for Lob in commerce cherry !
  to be about 1 ( ) and he is of the removal above of the
Stoder size * Day o play of Stoder lawer of Approva
 SN_PRICE - IN Electric of Street or social the SN for
  Learning race * 18.3
I book prosesses
"wid" + no roader, range (Nicolan, State, Woods, State (N. 00 or larger to Milden
With Tigo (Wildows Proceedings on No. 1920) 1979; T. H. & Partier to Trader
www.enc.rendor.rendolescen.atec. Atedes.atec? (0.00 = 10.00 to a consti-
for a salebreal (versit adject to), a same about
 def less-un/browns, pargets, horseld
     impacts, cargede are been list of incopers.
      returns the Local gradients on model parameters, and last misden state
    tof-cl = ex.opeltered
     Jack - a
        safet a se sersa (lasera aler, a)) a carele to a of a consequence of
        ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh historistical)
         pa[t] = ap.exp[ya]t]t > ex.exe(ap.exp[ya]t]t) + probabilization for more observed.
        tex-to-up-higher place; the place of the production from
      a majorant pass: compare gracients gody; mechanics
state, state, state, only a sp. state, likelikolj, og serve, likelikolj, og serve, likelikolj
      con, dby " rg.dwroe like(bb), np.merce like(by)
      chorac was owns. Mississists
      for the few mediarespectancepatry():
        do + mo.con/(soft)
        Astronomically 271 Programs who is
        den in m. desiry, hills. it
        di + np. doctety. . . dy9 + direct + betrarce ista n
         thow will a health happy had a hadron transported and real manage
        dry or ryme
        charle on the charles are profit I TI
        0/95 -- 10.000 05784. 11(1-1).71
        dieset - epicatout T, these;
     for spares to joint, sent, dety, dot; sheld
        op stigetheren. -6, 5, our-decree a city to account contribut gradients
      return took, dean, darm, darry, ear, day, bullestancetts til
    combined to the property of the property from the shoot
    S. F. SELENTONTHOUSE SIZE, 37 (
     tors a III
      is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

      a national (soletation 1);
april 12
       SAME RESIDENCE SAME
  MET, NOW FITS INTO LIGHTEN. ME, DOTEL LIGHTEN A RESET AND LIGHT OF ACRES
 senser_loss a lag logic strend_structure_loss to lost or interior white recen
    If 0-eeq leggs-1 = 10 lister 0 0 = 0
morey = 00.0010010010010.001, 1 1 = 1000 = 00.0007
      tigues - fator, or, to pay the entire to the Broatenay, he gottle
    respect from to supply for ships delighed grown to growing
      complete to complete process of managers, 1970.

Oil - 1 years of the size (and the second contracts).
      grade ..... who will be to the a
    here, there will place the size terms of the highest compute the contradiction of the contradictions of the c
    large a passive at print later by Later with a lar, assets later - print discovers
     for seven directs, we in significant tile, was try tell
                                                 THE MODE WAY, DOG TO VILL
      person in the rating ratio following in appropriate and a minute solution
```

#### Loss function

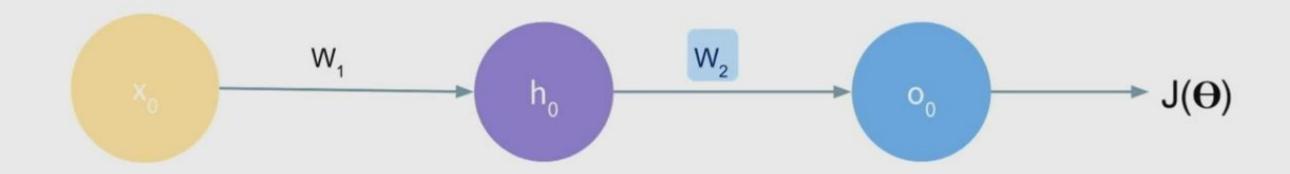
Backward pass

- forward pass (compute loss)
- backward pass (compute param gradient)

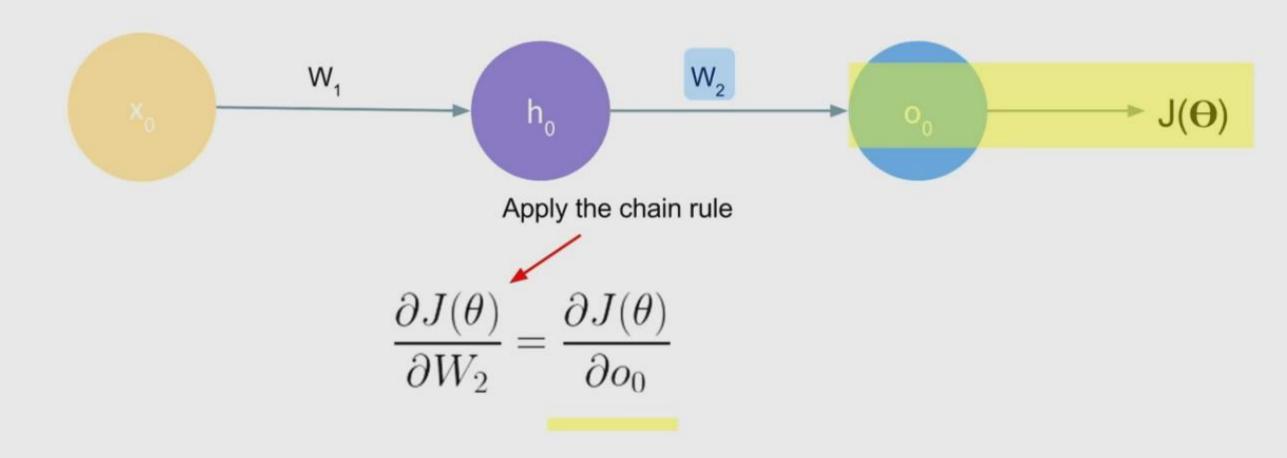
```
lossFun(inputs, targets, hprev):
       inputs, targets are both list of integers.
29
       hprev is Hx1 array of initial hidden state
31
       returns the loss, gradients on model parameters, and last hidden state
       xs, hs, ys, ps = {}, {}, {}, {}
      hs[-1] = np.copy(hprev)
34
       loss = 0
       # forward pass
       for t in xrange(len(inputs)):
        xs[t] = np.zeros((vocab_size,1)) # encode in 1-of-k representation
        xs[t][inputs[t]] = 1
        hs[t] = np.tanh(np.dot(Wxh, xs[t]) + np.dot(Whh, hs[t-1]) + bh) # hidden state
        ys[t] = np.dot(Why, hs[t]) + by # unnormalized log probabilities for next chars
42
         ps[t] = np.exp(ys[t]) / np.sum(np.exp(ys[t])) # probabilities for next chars
         loss += -np.log(ps[t][targets[t],0]) # softmax (cross-entropy loss)
24
       # backward pass: compute gradients going backwards
       dWxh, dWhh, dWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
       dbh, dby = np.zeros like(bh), np.zeros like(by)
       dhnext = np.zeros_like(hs[0])
       for t in reversed(xrange(len(inputs))):
         dy = np.copy(ps[t])
         dy[targets[t]] -= 1 # backprop into y
         dWhy += np.dot(dy, hs[t].T)
        dby += dy
         dh = np.dot(Why.T, dy) + dhnext # backprop into h
54
         dhraw = (1 - hs[t] * hs[t]) * dh # backprop through tanh nonlinearity
         dbh += dhraw
         dWxh += np.dot(dhraw, xs[t].T)
         dWhh += np.dot(dhraw, hs[t-1].T)
        dhnext = np.dot(Whh.T, dhraw)
       for dparam in [dwxh, dwhh, dwhy, dbh, dby]:
```

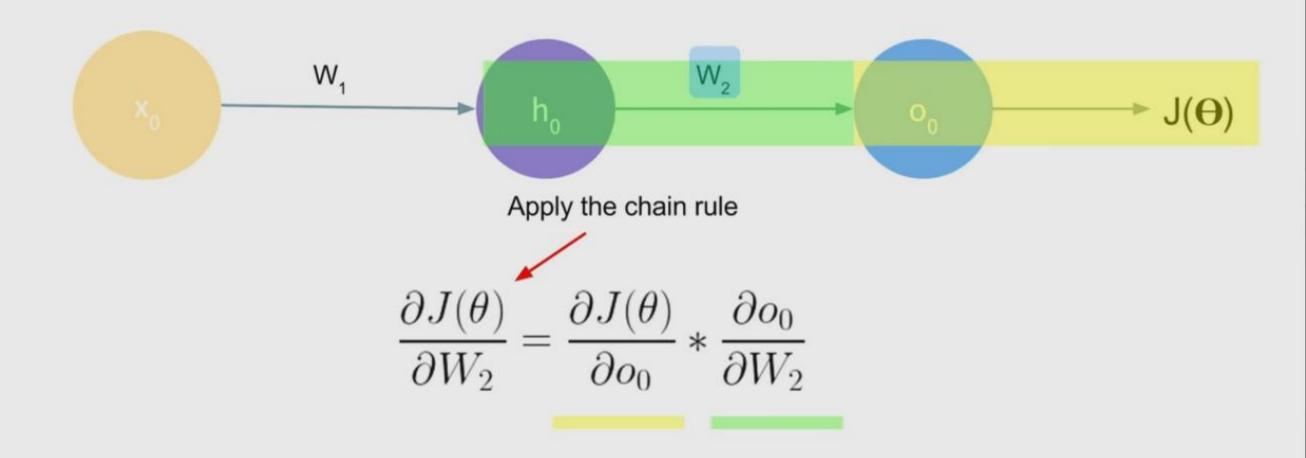
np.clip(dparam, -5, 5, out=dparam) # clip to mitigate exploding gradients

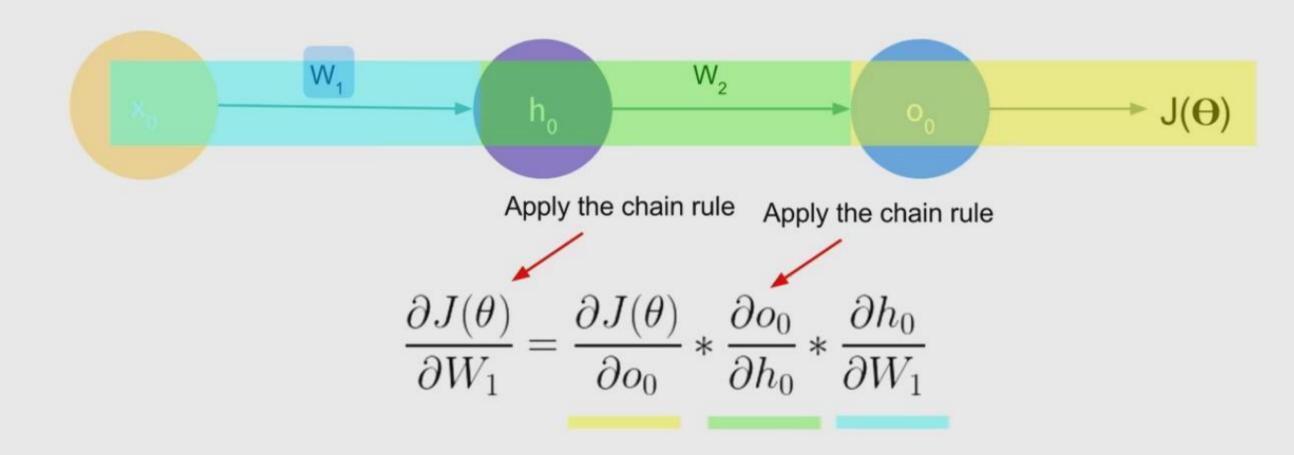
return loss, dWxh, dWhh, dWhy, dbh, dby, hs[len(inputs)-1]



$$\frac{\partial J(\theta)}{\partial W_2} =$$

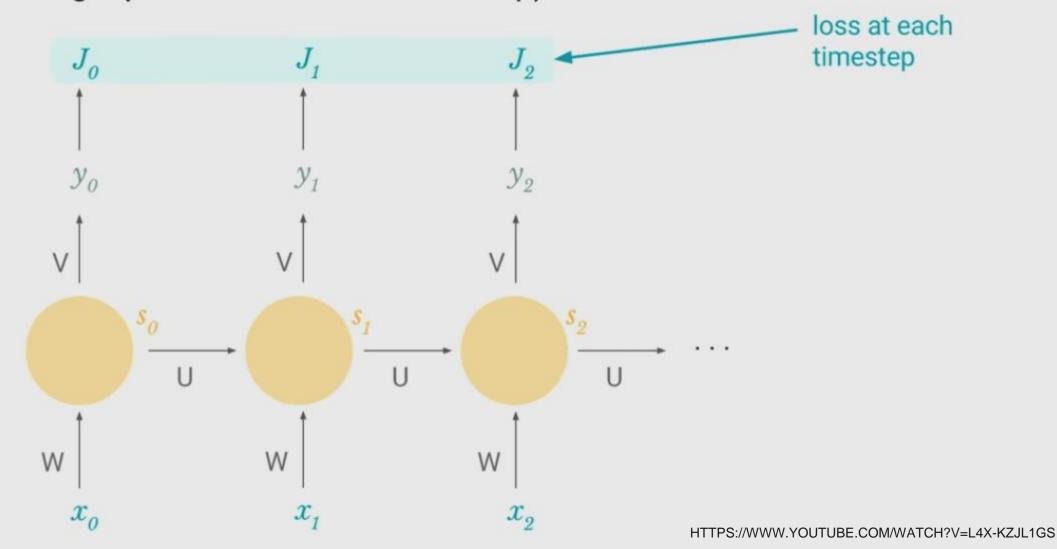




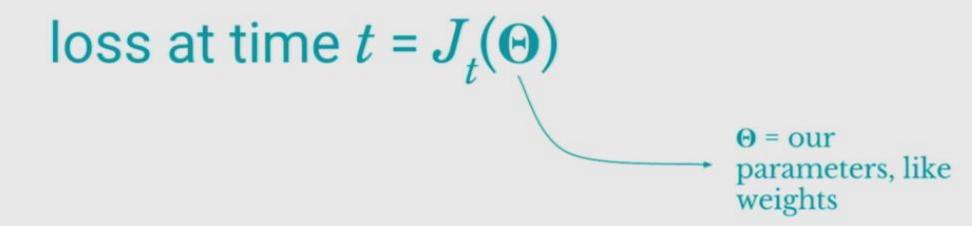


### we have a loss at each timestep:

(since we're making a prediction at each timestep)



### we **sum the losses** across time:



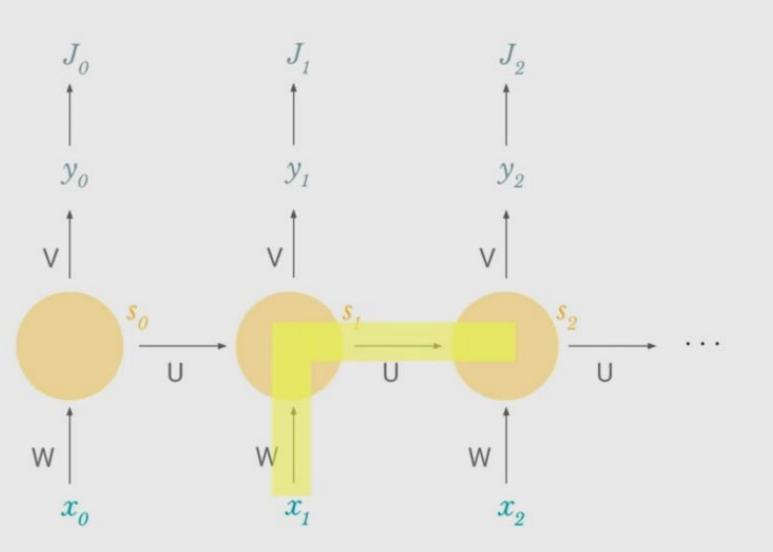
total loss = 
$$J(\mathbf{\Theta}) = \sum_{t} J_{t}(\mathbf{\Theta})$$

### what are our gradients?

## we sum gradients across time for each parameter P:

$$\frac{\partial J}{\partial P} = \sum_{t} \frac{\partial J_{t}}{\partial P}$$

### let's try it out for W with the chain rule:



$$\frac{\partial J}{\partial W} = \sum_{t} \frac{\partial J_{t}}{\partial W}$$

so let's take a single timestep t:

$$\frac{\partial J_2}{\partial W} = \frac{\partial J_2}{\partial y_2} \frac{\partial y_2}{\partial s_2} \frac{\partial s_2}{\partial W}$$

but wait...

$$s_2 = tanh(Us_1 + Wx_2)$$

 $s_1$  also depends on W so we can't just treat  $\frac{\partial s_2}{\partial W}$  as a constant!

# backpropagation through time:

$$\frac{\partial J_2}{\partial W} = \sum_{k=0}^{2} \frac{\partial J_2}{\partial y_2} \frac{\partial y_2}{\partial s_2} \frac{\partial s_2}{\partial s_k} \frac{\partial s_k}{\partial W}$$

Contributions of W in previous timesteps to the error at timestep t

# problem: vanishing gradient

$$\frac{\partial J_2}{\partial W} = \sum_{k=0}^{2} \frac{\partial J_2}{\partial y_2} \frac{\partial y_2}{\partial s_2} \frac{\partial s_2}{\partial s_k} \frac{\partial s_k}{\partial W}$$

$$\downarrow^{y_0} \qquad \downarrow^{y_1} \qquad \downarrow^{y_2} \qquad \downarrow^{s_1} \qquad \downarrow^{s_2} \qquad \text{at } k = 0: \qquad \frac{\partial s_2}{\partial s_0} = \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial s_0}$$

# problem: vanishing gradient

$$\frac{\partial J_2}{\partial W} = \sum_{k=0}^{2} \frac{\partial J_2}{\partial y_2} \frac{\partial y_2}{\partial s_2} \frac{\partial s_2}{\partial s_k} \frac{\partial s_k}{\partial W}$$

$$\downarrow^{y_0} \qquad \downarrow^{y_1} \qquad \downarrow^{y_2} \qquad \downarrow^{s_1} \qquad \downarrow^{s_2} \qquad \text{at } k = 0: \qquad \frac{\partial s_2}{\partial s_0} = \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial s_0}$$

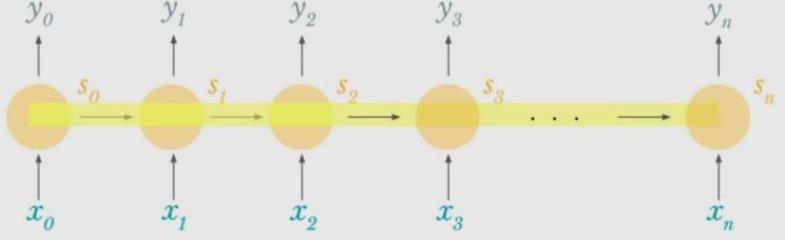
# problem: vanishing gradient

$$\frac{\partial J_n}{\partial W} = \sum_{k=0}^n \frac{\partial J_n}{\partial y_n} \frac{\partial y_n}{\partial s_n} \frac{\partial s_n}{\partial s_k} \frac{\partial s_k}{\partial W}$$

at k=0

$$\frac{\partial s_n}{\partial s_{n-1}} \frac{\partial s_{n-1}}{\partial s_{n-2}} \cdot \cdot \cdot \frac{\partial s_3}{\partial s_2} \frac{\partial s_2}{\partial s_1} \frac{\partial s_1}{\partial s_0}$$

as the gap between timesteps gets bigger, this product gets longer and longer!



```
min-char-rnn.py gist
                                                                                                                                                                                                                                                                                                                                                                             # backward pass: compute gradients going backwards
                                                                                                                                                                                                                                                                                                                                                                             dWxh, dWhh, dWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
                 Go through our sequence in reverse as
                                                                                                                                                                                                                                                                                                                                                                             dbh, dby = np.zeros_like(bh), np.zeros_like(by)
                  we back up the gradients
                                                                                                                                                                                                                                                                                                                                                                             dhnext = np.zeros_like(hs[0])
                                                                                                                                                                                                                                                                                                                                                                              for t in reversed(xrange(len(inputs))):
        to be about the first the residence of owner of
                                                                                                                                            gradient of the output
     Strates size Tipe - size of States lawer of Asiatan
                                                                                                                                                                                                                                                                                                                                                                                           dy = np.copy(ps[t])
       Learning rate * 18 3
                                                                                                                                                                                                                                                                                                                                                                                           dy[targets[t]] -= 1 # backprop into y
      with a military remark to the contract of the 
     WE'T SO ORGANIZACIONES CONTRACTO AND A TORREST AND A TORRE
                                                                                                                                            gradient of the W_{hy}
     www.w.mo.rondon.rumonlaneum.atec. httden.atec?/d.oz.w.
                                                                                                                                                                                                                                                                                                                                                                                            dWhy += np.dot(dy, hs[t].T)
     by a malebrary (Weepb. adject 10) is content about
      dof liver-un(inguts, sargets: horsy):
                                                                                                                                                                                                                                                                                                                                                                                            dby += dy
                                                                                                                                            gradient of the b_{
u}
         impacts, cargede are been list of incopers.
          returns the local gradients or model parameters, and last
                                                                                                                                                                                                                                                                                                                                                                                            dh = np.dot(Why.T, dy) + dhnext # backprop into h
         tof-tf - ex.comitment
                                                                                                                                                                                                                                                                                                                                                                                            dhraw = (1 - hs[t] * hs[t]) * dh # backprop through tanh nonlinearity
          Jack - a
                                                                                                                                            gradient of the h_t
             safet = w. merosi(lescent alar, a)) = macris
                                                                                                                                                                                                                                                                                                                                                                                            dbh += dhraw
              ha[t] = sp. tann(sp. documn, <math>sa[t]) - np.s
              yell (I - middel lifes, led ill a he a
                                                                                                            gradient of the tanh(h)_t
              ps[t] = np.com(ys[t]) / sp.sum(np.com
                                                                                                                                                                                                                                                                                                                                                                                            dWxh += np.dot(dhraw, xs[t].T)
              tes to an importificant dil.
           rively state, and a vagine
           can, day " relative likelihet, np merce like
                                                                                                                                                                                                                                                                                                                                                                                           dWhh += np.dot(dhraw, hs[t-1].T)
           chores was ween likeliabile.
            for the American promption (copate) is
                                                                                                                                            gradient of the b_h
             do + mo.con/(soft)
                                                                                                                                                                                                                                                                                                                                                                                           dhnext = np.dot(Whh.T, dhraw)
              deliamentally of the accuracy where
              duty in to design highlin
             di e so dortale. L. del e discor e l'amorce forcin
                                                                                                                                                                                                                                                                                                                                                                             for dparam in [dWxh, dWhh, dWhy, dbh, dby]:
              thow off the best filter of the best of th
                                                                                                                                            gradient of the W_{xh}
             dis -- swan
              charle on the charles are and (1.7)
              duan or aplication and training
                                                                                                                                                                                                                                                                                                                                                                                           np.clip(dparam, -5, 5, out=dparam) # clip to mitigate exploding gradients
             dieset - epicatout. T, these;
         for spares as [seen, seen, dayly don, doy]; 
epichardparen, -6, 5, cereberrari e chia
                                                                                                                                            gradient of the W_{hh}
                                                                                                                                                                                                                                                                                                                                                                             return loss, dWxh, dWhh, dWhy, dbh, dby, hs[len(inputs)-1]
         combined companies of terrorise from the second
           however, and work to be and below for 'that also may
         > 7 No. perpeny vocato espe. (c.)
         Sincet. Di + 5.
           b Trailed to and Mar. At + sp. solomb, Mr + Mr.

    * replacement () * by
    * replacement () * replacement () ()
    * replacement (replacement () () ()
    * replacement (replacement () () ()

           a manager (makes to 1);
            SAME, MESSAGES SAVE
       MET, MON PITA JETOS JERNESNI MELDOTEL LIGHTANI V ROBERT NAVIALISE FOR HOSETAN
      server_loss a vag lagra alterna_condition_lossible in the internet while there
         AT CHARG SENSET OF AN IDSTALL OF STATE 
          tigues - false, or, higher this in the deal process, hope of t
          respect from to origin; for all to delegatives in generally
```

simple in simple year in accept, and the "year in the last the second acts when we have a section."

for server, directly over its significant falls, was too tagli-

News, chart, child, child, child, child, have "tenders (report, larger), harves associations a second later " 2.000 - later " 5.001 have " 6.001 have " 10.001 have " 2.000 - later " 5.001 have " 10.000 have " 5.000 have " 5.00

person in the original Colored Colored (see 1964) is selected united.

Take are any, on act

$$\frac{\partial L}{\partial y} = \sum_{t} \frac{\partial L_{t}}{\partial y_{t}}$$

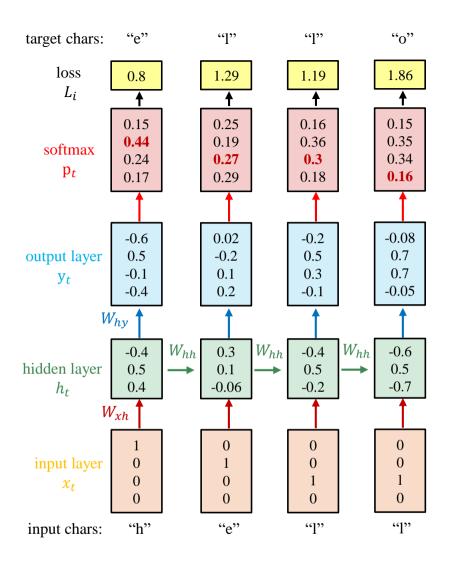
$$\frac{\partial L_t}{\partial y_t} = output - target$$

$$\frac{\partial L_4}{\partial y_4} = \begin{bmatrix} 0.15 \\ 0.35 \\ 0.34 \\ 0.16 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.15 \\ 0.35 \\ 0.34 \\ -0.84 \end{bmatrix} \qquad \frac{\partial L_3}{\partial y_3} = \begin{bmatrix} 0.16 \\ 0.36 \\ 0.3 \\ 0.18 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.16 \\ 0.36 \\ -0.7 \\ 0.18 \end{bmatrix}$$

$$\frac{\partial L_3}{\partial y_3} = \begin{bmatrix} 0.16 \\ 0.36 \\ 0.3 \\ 0.18 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.16 \\ 0.36 \\ -0.7 \\ 0.18 \end{bmatrix}$$

$$\frac{\partial L_2}{\partial y_2} = \begin{bmatrix} 0.25 \\ 0.19 \\ 0.27 \\ 0.29 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.15 \\ 0.19 \\ -0.73 \\ 0.29 \end{bmatrix} \qquad \frac{\partial L_1}{\partial y_1} = \begin{bmatrix} 0.15 \\ 0.44 \\ 0.24 \\ 0.17 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.15 \\ -0.56 \\ 0.24 \\ 0.17 \end{bmatrix}$$

$$\frac{\partial L_1}{\partial y_1} = \begin{bmatrix} 0.15\\ 0.44\\ 0.24\\ 0.17 \end{bmatrix} - \begin{bmatrix} 0\\1\\0\\0 \end{bmatrix} = \begin{bmatrix} 0.15\\ -0.56\\ 0.24\\ 0.17 \end{bmatrix}$$



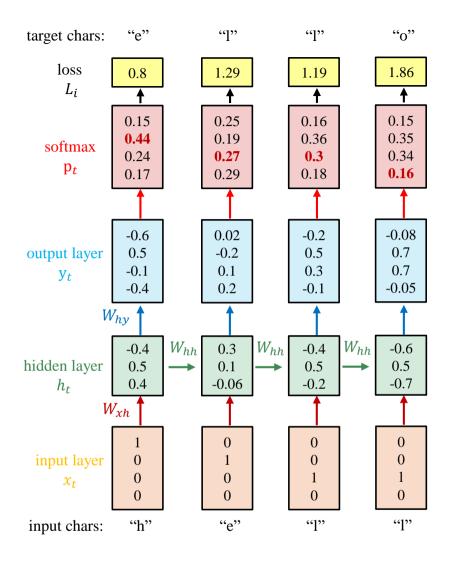
$$\frac{\partial L}{\partial W_{hy}} = \sum_{t} \frac{\partial L}{\partial y_{t}} \frac{\partial y_{t}}{\partial W_{hy}}$$

$$h_{t}$$

$$\frac{\partial L}{\partial W_{hy}} = \frac{\partial L}{\partial y_4} h_4 + \frac{\partial L}{\partial y_3} h_3 + \frac{\partial L}{\partial y_2} h_2 + \frac{\partial L}{\partial y_1} h_1$$

$$\frac{\partial L}{\partial W_{hy}} = \begin{bmatrix} 0.15 \\ 0.35 \\ 0.34 \\ -0.84 \end{bmatrix} \begin{bmatrix} -0.6 & 0.5 & -0.7 \end{bmatrix} + \begin{bmatrix} 0.16 \\ 0.36 \\ -0.7 \\ 0.18 \end{bmatrix} \begin{bmatrix} -0.4 & 0.5 & -0.2 \end{bmatrix}$$

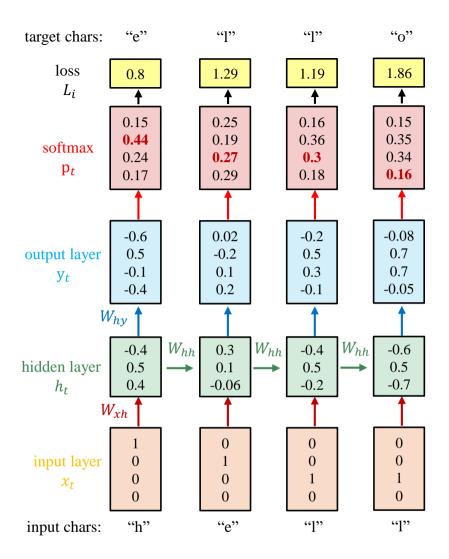
$$+\begin{bmatrix} 0.15 \\ 0.19 \\ -0.73 \\ 0.29 \end{bmatrix} [0.3 \quad 0.1 \quad -0.06] + \begin{bmatrix} 0.15 \\ -0.56 \\ 0.24 \\ 0.17 \end{bmatrix} [-0.4 \quad 0.5 \quad 0.4] = \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ -0.06 & 0.1 & -0.5 \\ -0.2 & -0.1 & 0.05 \\ 0.5 & -0.2 & 0.6 \end{bmatrix}$$



$$\frac{\partial L}{\partial b_y} = \sum_{t} \frac{\partial L}{\partial y_t} \frac{\partial y_t}{\partial b_y}$$

$$\frac{\partial L}{\partial b_y} = \frac{\partial L}{\partial y_4} + \frac{\partial L}{\partial y_3} + \frac{\partial L}{\partial y_2} + \frac{\partial L}{\partial y_1}$$

$$\frac{\partial L}{\partial b_y} = \begin{bmatrix} 0.15 \\ 0.35 \\ 0.34 \\ -0.84 \end{bmatrix} + \begin{bmatrix} 0.16 \\ 0.36 \\ -0.7 \\ 0.18 \end{bmatrix} + \begin{bmatrix} 0.15 \\ 0.19 \\ -0.73 \\ 0.29 \end{bmatrix} + \begin{bmatrix} 0.15 \\ -0.56 \\ 0.24 \\ 0.17 \end{bmatrix} = \begin{bmatrix} 0.6 \\ 0.3 \\ -0.8 \\ -0.1 \end{bmatrix}$$

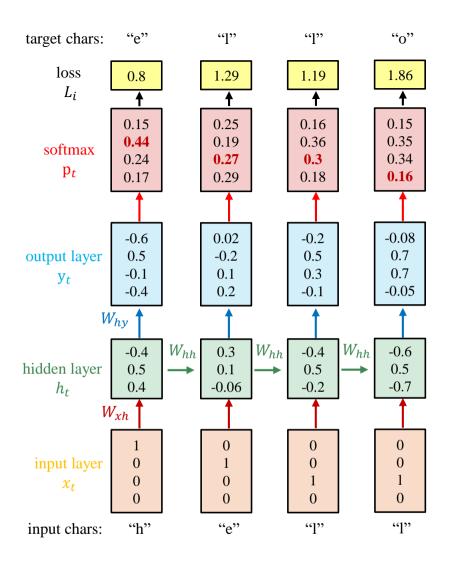


$$\frac{\partial L}{\partial b_{h}} = \sum_{t} \frac{\partial L}{\partial h_{t}} \frac{\partial h_{t}}{\partial b_{h}} \qquad \frac{\partial L}{\partial h_{1}} = \frac{\partial h_{k}}{\partial h_{k-1}} \frac{\partial h_{k-1}}{\partial h_{k-2}} \frac{\partial h_{k-2}}{\partial h_{k-3}} \dots \frac{\partial h_{2}}{\partial h_{1}}$$

$$(1 - h_{t}^{2}) * 1$$

$$\frac{\partial L}{\partial h_{1}} = \frac{\partial h_{k}}{\partial h_{k-1}} \frac{\partial h_{k-1}}{\partial h_{k-2}} \frac{\partial h_{k-2}}{\partial h_{k-3}} ... \frac{\partial h_{2}}{\partial h_{1}}$$

$$\frac{\partial L}{\partial b_h} = \begin{bmatrix} 0.06 \\ -0.5 \\ -0.1 \end{bmatrix}$$



$$\frac{\partial L}{\partial W_{hh}} = \sum_{t} \frac{\partial L_{t}}{\partial W_{hh}}$$

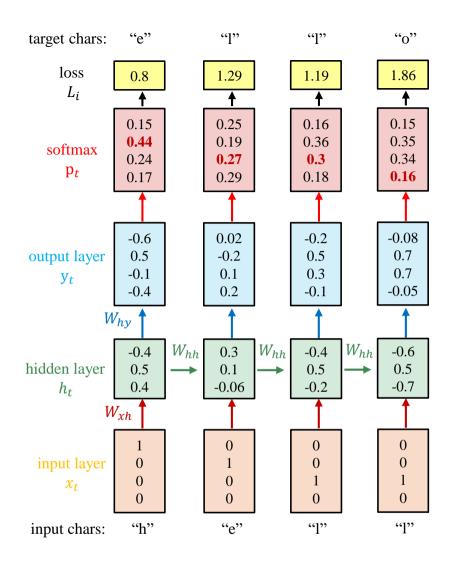
$$\frac{\partial L_t}{\partial W_{hh}} = \sum_{k=1}^t \frac{\partial L_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W_{hh}}$$

$$(1 - h_t^2) * h_{t-1}$$

$$at k = 1 \quad \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_2}{\partial h_1}$$

$$at k = 2 \quad \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_3}{\partial h_2}$$

$$\dots$$



$$\frac{\partial L}{\partial W_{hh}} = \begin{bmatrix} 0.09 & -0.1 & 0.1\\ 0.1 & -0.1 & -0.1\\ -0.05 & 0.1 & 0.07 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_{xh}} = \sum_{t} \frac{\partial L_t}{\partial W_{xh}}$$

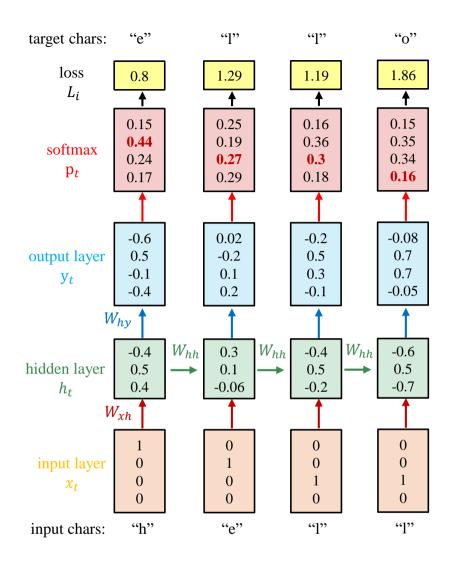
$$\frac{\partial L_t}{\partial W_{xh}} = \sum_{k=1}^t \frac{\partial L_t}{\partial y_t} \frac{\partial y_t}{\partial h_t} \frac{\partial h_t}{\partial h_k} \frac{\partial h_k}{\partial W_{xh}}$$

$$(1 - h_t^2) * x_t$$

$$at k = 1 \quad \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_2}{\partial h_1}$$

$$at k = 2 \quad \frac{\partial h_t}{\partial h_{t-1}} \frac{\partial h_{t-1}}{\partial h_{t-2}} \dots \frac{\partial h_3}{\partial h_2}$$

$$\dots$$



$$\frac{\partial L}{\partial W_{xh}} = \begin{bmatrix} 0.5 & 0.04 & -0.5 & 0 \\ -0.2 & -0.3 & 0.1 & 0 \\ -0.4 & 0.1 & 0.1 & 0 \end{bmatrix}$$

#### min-char-rnn.py gist

```
No sed country-level Worlds MA solel, Written by habity Repailsy (Brognitty)
www.doense
DIRECT PURPS AS FA
14140406-1412
                                                                                                                                                                                                                                                                                                    81
 does a report tracturer; a "financial a woode on a testing hair cost file.
 Chart - Theretake Backet t
dona stor, excessator e limitadat, lesteneral
government and the properties, he entire. In printage on work your
 char to ic = 1 chil for Lob in commerce cherry !
 to be about 1 ( ) and he is of the removal above of the
Sictor size * De o size of Middle layer of regrand
 SN_PRICE - IN Electric of Street or social the 200 for
 Learning race * 18 1
I book prosesses
 with a new roader, ramen@tolerunder; wood_vizeCF0.00 is input to hillion.
WW TIES, GROOM, PRINCIPLISHED, KIZH, CHARGES, KIZHOTE, ATTEMPT OF THE REC.
www.e. ac render, reports seem after. Atteden, advertig on a minder on compart
by a salphray("Vecub adpo. '19) a common stars
                                                                                                                                                                                                                                                                                                    87
 def lease-in/trauta, parages, horovit-
    inputs, targets are beth list of incapers.
                                                                                                                                                                                                                                                                                                    88
     returns the Local gradients on eadel parameters, and last misden state
                                                                                                                                                                                                                                                                                                    89
     se. W. Se. PE " Co. Cl. (D. C)
    tof-cl = ex.opeltered
    Jack - a
     far I be arresected reaction.
       safet a se sersa (lasera ateria) i a carrie te a ef-a concentrarias
                                                                                                                                                                                                                                                                                                    91
       ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh histor, state)
        will be madefully, but if a be a communication or an absorbit on the analysis.
        pa[t] = ap.exp[ya]t]t \ge ap.ava(ap.exp[ya]t]t) + brobabilities for race chara-
       tion to an important property of the product of the
     e papieras para: compute practicata, 20030 percenta-
      each, side, side a open usual harbord, open oschlarpitor, open oschlarpitori.
     con, dby " rg.dwroe like(bb), np merce like(by)
     chorac - wa were likelikabila
      for the developing and temperatures;
                                                                                                                                                                                                                                                                                                    94
       do + more gray(molth)
       detrangenetation of the secretary arms of
       detry on the decisy, builting
       di = np.doctahy... dy9 + dimos + betrorog into n
       those sign and the second second second second second second
       dry - rym
       charle on the charles are and (1.7)
        duals on application, strict of all
                                                                                                                                                                                                                                                                                                    97
       discut - epicatous T, those;
    for spares on [seek, sent, dety, dot. dog]:
       op stigetheres, 45 %, orreduced a city to attach confedent and tiers
    return seek, dean, dam, early, ean, day, begandpackets) 13
   Det manufacts, seed his first
                                                                                                                                                                                                                                                                                                    99
    combined to the property of the property from the shoot
    In the energy date, what is to smill be sent to "their him to be
   > 7 Nel Demonstraction delice. (Er)
Sincrellar) = 1.
face. a []
                                                                                                                                                                                                                                                                                               100
      is transactional balls of a sould be at the art.
                                                                                                                                                                                                                                                                                               101

    * replacement () * by
    * replacement () * replacement () ()
    * replacement (replacement () () ()
    * replacement (replacement () () ()

                                                                                                                                                                                                                                                                                               102
      a national (soletation 1);
april 12
      SAME, MESSAGES SAV
                                                                                                                                                                                                                                                                                               103
                                                                                                                                                                                                                                                                                               104
 MET, NOW FITE INFOR IZERVENT NEIDERS LIGHTEN A TORREST
   second_loss a log logic stylend_stantifeed_losgic block is from a
               update parameters (W_{hy}, W_{hh}, W_{xh}, b_y, b_h)
               with Adagrad
                                                                                                                                                                                                                                                                                               109
    into a passes at print there so, while with a few assets into a print property
                                                                                                                                                                                                                                                                                               110
     he seems district over its significant bits, was to, og-
                                                                                                                                                                                                                                                                                               111
                                            TIME THE MAY THE MY IT
      person in the original Colored Colored (see a sea) is selected as a
                                                                                                                                                                                                                                                                                               112
```

## Main loop

```
n, p = 0, 0
mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
while True:
  # prepare inputs (we're sweeping from left to right in steps seg length long)
  if p+seq length+1 >= len(data) or n == 0:
    hprev = np.zeros((hidden size,1)) # reset RNN memory
    p = 0 # go from start of data
  inputs = [char to ix[ch] for ch in data[p:p+seq length]]
  targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
  # sample from the model now and then
  if n % 100 == 0:
    sample_ix = sample(hprev, inputs[0], 200)
    txt = ''.join(ix_to_char[ix] for ix in sample_ix)
    print '----\n %s \n----' % (txt. )
  # forward seq_length characters through the net and fetch gradient
  loss, dWxh, dWhh, dWhy, dbh, dby, hprev = lossFun(inputs, targets, hprev)
  smooth loss = smooth loss * 0.999 + loss * 0.001
  if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) # print progress
  # perform parameter update with Adagrad
  for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
                                 [dWxh, dWhh, dWhy, dbh, dby],
                                 [mWxh, mWhh, mWhy, mbh, mby]):
    mem += dparam * dparam
    param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
  p += seq_length # move data pointer
```

n += 1 # iteration counter

Adagrad is a more efficient technique where the learning rate ( $\alpha$ ) are getting smaller during the training.

SGD 
$$\theta_{t+1} = \theta_t - \alpha * \frac{\partial L}{\partial \theta_t}$$

Adagrad 
$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t}$$
  $\epsilon = 1e-8$ 

#### for iteration 1:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t} \qquad \epsilon = 1e-8$$

$$\alpha = 0.1$$

$$mW_{xh} += \begin{bmatrix} 0.2 & 0.002 & 0.2 & 0 \\ 0.07 & 0.1 & 0.01 & 0 \\ 0.1 & 0.03 & 0.01 & 0 \end{bmatrix}$$

$$W_{xh} = \begin{bmatrix} -0.5 & 0.4 & -0.2 & 0.3 \\ 0.5 & -0.3 & 0.6 & -0.8 \\ 0.4 & 0.4 & -0.3 & -0.8 \end{bmatrix} - \frac{0.1}{\sqrt{0.000000008 + mW_{xh}}} \begin{bmatrix} 0.5 & 0.04 & -0.5 & 0 \\ -0.2 & -0.3 & 0.1 & 0 \\ -0.4 & 0.1 & 0.1 & 0 \end{bmatrix} = \begin{bmatrix} -0.6 & 0.3 & -0.1 & 0.3 \\ 0.6 & -0.2 & 0.5 & -0.8 \\ 0.5 & 0.3 & -0.4 & -0.8 \end{bmatrix}$$

#### for iteration 1:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t} \qquad \epsilon = 1e-8$$

$$\alpha = 0.1$$

$$mW_{hy} += \begin{bmatrix} 0.01 & 0.3 & 0.2 \\ 0.4 & 0.09 & 0.06 \\ 0.01 & 0.1 & 0.5 \\ 0.2 & 0.002 & 0.1 \end{bmatrix}$$

$$W_{hy} = \begin{bmatrix} 0.1 & -0.6 & -0.5 \\ -0.7 & 0.3 & -0.08 \\ 0.1 & 0.4 & -0.7 \\ 0.5 & 0.05 & -0.4 \end{bmatrix} - \frac{0.1}{\sqrt{0.000000008 + mW_{hy}}} \begin{bmatrix} 0.1 & 0.2 & -0.1 \\ -0.06 & 0.1 & -0.5 \\ -0.2 & -0.1 & 0.05 \\ 0.5 & -0.2 & 0.6 \end{bmatrix} = \begin{bmatrix} 0.2 & -0.7 & -0.4 \\ -0.6 & 0.2 & 0.01 \\ 0.2 & 0.5 & -0.8 \\ 0.4 & 0.1 & -0.5 \end{bmatrix}$$

#### for iteration 1:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t} \qquad \epsilon = 1e-8$$

$$\alpha = 0.1$$

$$mW_{hh} += \begin{bmatrix} 0.08 & 0.03 & 0.01 \\ 0.01 & 0.01 & 0.02 \\ 0.002 & 0.01 & 0.005 \end{bmatrix}$$

$$W_{hh} = \begin{bmatrix} -0.1 & -0.8 & 0.7 \\ -0.08 & 0.2 & 0.7 \\ 0.5 & -0.5 & 0.01 \end{bmatrix} - \frac{0.1}{\sqrt{0.000000008 + mW_{hh}}} \begin{bmatrix} 0.09 & -0.1 & 0.1 \\ 0.1 & -0.1 & -0.1 \\ -0.05 & 0.1 & 0.07 \end{bmatrix} = \begin{bmatrix} -0.2 & -0.7 & 0.6 \\ -0.1 & 0.3 & 0.8 \\ 0.6 & -0.6 & -0.08 \end{bmatrix}$$

#### for iteration 1:

$$(b_h)^2$$

$$mb_h += \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$b_h = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \frac{0.1}{\sqrt{0.000000008 + mb_h}} \begin{bmatrix} 0.06 \\ -0.5 \\ -0.1 \end{bmatrix} = \begin{bmatrix} -0.09 \\ 0.09 \\ 0.09 \end{bmatrix}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t} \qquad \epsilon = 1e-8$$

$$\alpha = 0.1$$

#### for iteration 1:

$$mb_y += \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$b_{y} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \frac{0.1}{\sqrt{0.000000008 + mb_{y}}} \begin{bmatrix} 0.6 \\ 0.3 \\ -0.8 \\ -0.1 \end{bmatrix} = \begin{bmatrix} -0.09 \\ -0.09 \\ 0.09 \\ 0.09 \end{bmatrix}$$

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\sqrt{\epsilon + \sum_j \left(\frac{\partial L}{\partial \theta_j}\right)^2}} * \frac{\partial L}{\partial \theta_t} \qquad \epsilon = 1e-8$$

$$\alpha = 0.1$$

#### min-char-rnn.py gist

```
For each class with a level West Har Mid. model , No of the by Judicy Regardly (Blangadly)
www.doense
DIRECT PURPS AS FA
14140406-1412
 does a report tracturer; a "financial a woode on a testing hair cost file.
  Chart - Theretake Backet t
done stor, exem stor a leadannt, leatehoral
government and the properties, he entire. In printage on work your
  char to ic = 1 chil for Lob in commerce cherry !
  to be about 1 ( ) and he is of the removal above of the
Stoder spe Tow - size of Stoden lawer of courses
 SN_PRICE - IN Electric of Street or social the 200 for
  Learning race * 18 1
I book prosesses
 with a new roader, ramen@tolerunder; wood_vizeCF0.00 is input to hillion.
WW TIES, GROOM, PRINCIPLISHED, KIZH, CHARGES, KIZHOTE, ATTEMPT OF THE REC.
www.e. ac render, reports seem after. Atteden, advertig on a minder on compart
by a salphray("Vecub adpo. '19) a common stars
 def lease-in/trauta, parages, horovit-
      inputs, targets are beth list of incapers.
        returns the Local gradients on eadel parameters, and last misden state
        se. W. Se. PE " Co. Cl. (D. C)
      tof-cl = ex.opeltered
       Jack - a
        fac t in anesected/mostrati-
          safet a se sersa (lasera ateria) i a carrie te a ef-a concentrarias
           ha[t] = sp. tenh(sp. decrees, sa[t]) - np.cen(whb, <math>ha[t-1]) - bh(sh histor, state)
            will be madefully, but if a be a communication or an absorbit on the analysis.
            pa[t] = ap.exp[ya]t]t \ge ap.ava(ap.exp[ya]t]t) + brobabilities for race chara-
          tion to an important property of the product of the
        e papieras para: compute practicata, 20030 percenta-
         each, side, side a open usual harbord, open oschlarpitor, open oschlarpitori.
        con, dby " rg.dwroe like(bb), np merce like(by)
        chorac - wa were likelikabila
         for the developing and temperatures;
           do + more gray(molth)
           deliamentally and in some or with a
           detry on the decisy, builting
          di a se derigio i del a discreti a batterior fatte in
           those sign and the second second second second second second
           dry - rym
           charle on the charles are and (1.7)
            duals on application, strict of all
           discut - epicatous T, those;
       for spares on [seek, sent, dety, dot. dog]:
           op stigetheres, 45 %, orreduced a city to attach confedent and tiers
       return seek, dean, dam, early, ean, day, begandpackets) 13
  per admitted, seed the first
     combined an income of terrorise from the source
       I however, said, work to be and have the first time one
     > 7 Nel Demonstraction delice. (Er)
Sincrellar) = 1.
face. a []
         is transactional balls of a sould be at the art.

    * replacement () * ty
    * a replacement (replacement ())
    * a replacement (replacement ())
    * a replacement (replacement ())

         a national (soletation 1);
april 12
         SAME RESIDENCE SAME
  MET, NOW A TO LEGISLES HEADY NELDONES, LICEUSE A RESERVE ANGULAR THE ACRESS
  second_loss = log logic strend_scriptos_logic = loss of relevative while receive
     or oregularites - telligital of e = 0
     tigens - julia control of the control of pursuant trapect; 
respect to julia to supply for all in deling-disperse trapect().
        complete to desproy to accomply, 900
bit "yearlie to desproy to accompanie or)
must be up to the property of
     Then, Ching (Mary Mary Mary Street ) Streething Top (10), I compared the control face (10) of th
        he seems district over its significant bits, was to, og-
                                                                TIME THE THE THE TENTON
         person in the rating ratio 1 dynamic appropriate a recall in migration to
```

### Main loop

81

87

88

89

91

94

97

99

100

101

102

103

104

105

106

107

108

109

111

112

Test time

```
n, p = 0, 0
mWxh, mWhh, mWhy = np.zeros_like(Wxh), np.zeros_like(Whh), np.zeros_like(Why)
mbh, mby = np.zeros like(bh), np.zeros like(by) # memory variables for Adagrad
smooth_loss = -np.log(1.0/vocab_size)*seq_length # loss at iteration 0
while True:
  # prepare inputs (we're sweeping from left to right in steps seg length long)
  if p+seq length+1 >= len(data) or n == 0:
    hprev = np.zeros((hidden size,1)) # reset RNN memory
    p = 0 # go from start of data
  inputs = [char to ix[ch] for ch in data[p:p+seq length]]
  targets = [char to ix[ch] for ch in data[p+1:p+seq length+1]]
  # sample from the model now and then
  if n % 100 == 0:
    sample_ix = sample(hprev, inputs[0], 200)
    txt = ''.join(ix_to_char[ix] for ix in sample_ix)
    print '----\n %s \n----' % (txt. )
  # forward seq_length characters through the net and fetch gradient
  loss, dWxh, dWhh, dWhy, dbh, dby, hprev = lossFun(inputs, targets, hprev)
  smooth loss = smooth loss * 0.999 + loss * 0.001
  if n % 100 == 0: print 'iter %d, loss: %f' % (n, smooth_loss) # print progress
  # perform parameter update with Adagrad
  for param, dparam, mem in zip([Wxh, Whh, Why, bh, by],
                                 [dWxh, dWhh, dWhy, dbh, dby],
                                 [mWxh, mWhh, mWhy, mbh, mby]):
    mem += dparam * dparam
    param += -learning_rate * dparam / np.sqrt(mem + 1e-8) # adagrad update
  p += seq_length # move data pointer
  n += 1 # iteration counter
```

#### min-char-rnn.py gist

```
Noticed viscosity level Worlds Mill model, Written by Joshey Reports, (Brognitty)
www.doense
DIRECT PURPS AS FA
 does - specifications; or himself - woods as about glob over file
done stor, exem stor a leadannt, leatehoral
government and the properties, he entire. In printage on work your
 char. no. ic = 1 chil for L ch in commerce charge 1
 to boundary of cours his apply to resource at overall a
Miches size * Day - size of Miches layer of regrand
 sw_brager - Dr a carbon of binning property the 200 for
 Dearridge race * 18 1
I book prosesses
"wid" + no roader, range (Nicolan, State, Woods, State (N. 00 or larger to Milden
WW TIES, GROOM, PRINCIPLISHED, KIZH, CHARGES, KIZHOTE, ATTEMPT OF THE REC.
www.wo.rowdor.rowonlescop.aige. httdcs_nigetYd.oi = niddor.in.coppt
by a salphray("Vecub adpo. '19) a common stars
 def lease-in/traute, sargets, horovit-
      inputs, targets are beth list of incapers.
       returns the local gradients on eadel parameters, and last missen state
        se. W. Se. PE " Co. Cl. (D. C)
      tof-cl = ex.opeltered
       Jack - a
        fac t in anesected/mostrati-
           saft) - w. seros((socen_alan,a)) - excels in a-gf-2 representation
           \operatorname{du}[t] = \operatorname{sp.} \operatorname{min}(\operatorname{sp.} \operatorname{dosness}, \operatorname{xu}[t]) - \operatorname{sp.}\operatorname{dut}(\operatorname{wit}, \operatorname{hu}[t-1]) - \operatorname{ht}) + \operatorname{histor} \operatorname{min}
            will be a safet (May, 1941) if the a second control of an absorbing the treat when
             pa[t] = np.exp[ya]t]t \ge ex.exe(np.exp[ya]t]t) = probabilization for most often
           tion to an important property of the product of the
        e papieras para: compute practicata, 20030 percenta-
         each, side, side a open usual harbord, open out including open out the play.
        con, dby " rg.dwroe like(bb), np merce like(by)
        chorac - wa were likelikabila
         for the developing and temperatures;
            do + more gray(molth)
            detrangenetation of the secretary arms of
            den in m. desiry, hifti. it
           di = np.doctahy... dy9 + dimos + betrorog into n
            those sign and the second second second second second second
           dia -- cares
            charle on the charles are and (1.7)
            dean or sp. doc|cares. na(t-s).r)
            discut a special older, there;
      for spares in joint, sent, dety, don, doy;
            op adjudgement, 45, 5, department a clip to adjuste controlles analysis a
        return took, dean, dam, days, ean, day, better(basets) 13
     combined to the property of the courty from the shoot
    > 7 %.prostivous size. 17)
sizeri.in[ * 1
ter. + []
te ( to recognise)
       to it is complete.

It considers the conflicts of the scale (Mail, b) that ],

It considers to the term of the scale (Mail, b) that ],

It considers to the scale to the scale (Mail, b) that ],

It is a greatest to the scale (Mail, b) that ], page 100-1011.
         a national facial size of
       2400,00000000000
100,00 0000
  MET, NOW A THE LEGISLE CONTROL AND DESCRIPTION OF A PARTY AND ADDRESS TO A ADDRESS.
 where the \alpha -applies a constant of the confidence of the state of the constant while the \alpha
      transport acres. Delta sample from this could be observed built form.
     If 0-eeq leggs-1 = 10 lister 0 0 = 0
morey = 00.0010010010010.001, 1 1 = 1000 = 00.0007
       tigues - (pleas, oc. bis) of the entire transaction (CI.
      respect from to supply for ships deligned grown to growing
        complete to sent before a managing (80).

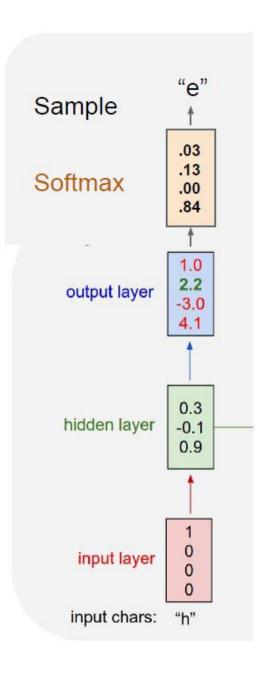
Let a "greative be also just the second contracts of

    Adaptic Section (Association association and Could professor
Sections, 1980). May, May, May, Inc. Inc. Section (Inc.) association in a country of the country 
       for seven direct, we in sighter, six, we, it, set-
                                                                   Times are may see act
         person of the original of the contract of the contract of the contract of the
      p. et gag Amight it note cata prairies
```

#### test-time

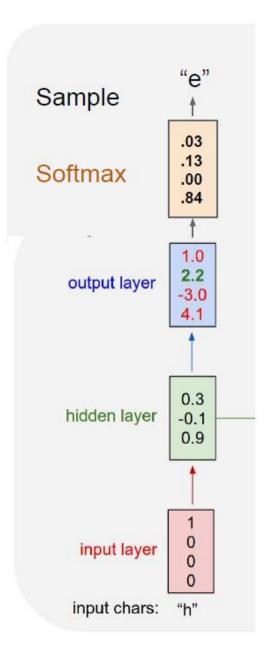
```
def sample(h, seed ix, n):
       11 11 11
64
       sample a sequence of integers from the model
       h is memory state, seed_ix is seed letter for first time step
66
       11 11 11
       x = np.zeros((vocab_size, 1))
      x[seed_ix] = 1
69
       ixes = []
70
       for t in xrange(n):
         h = np.tanh(np.dot(Wxh, x) + np.dot(Whh, h) + bh)
73
         y = np.dot(Why, h) + by
         p = np.exp(y) / np.sum(np.exp(y))
74
         ix = np.random.choice(range(vocab_size), p=p.ravel())
76
         x = np.zeros((vocab_size, 1))
         x[ix] = 1
         ixes.append(ix)
       return ixes
```

Vocabulary: [h,e,l,o]



Vocabulary: [h,e,l,o]

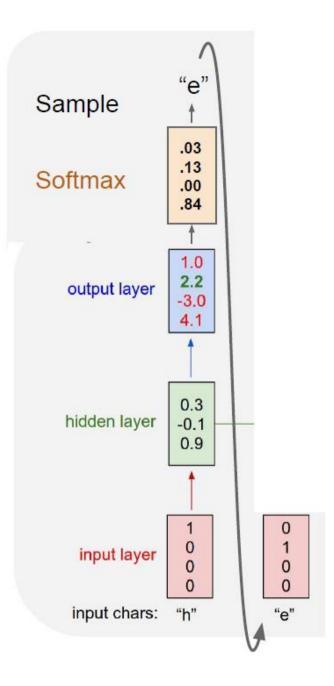
At test-time sample characters one at a time, feed back to model



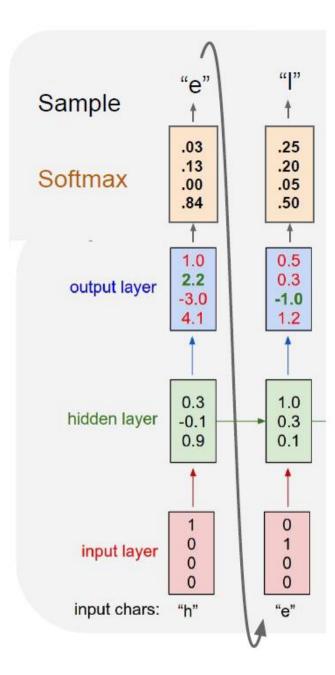
### Why might we sample?

It lets you get diversity from the model.

Vocabulary: [h,e,l,o]



Vocabulary: [h,e,l,o]



Vocabulary: [h,e,l,o]

