Boulder

Speech & Music; Synthesis

YouTube Playlist

Maziar Raissi

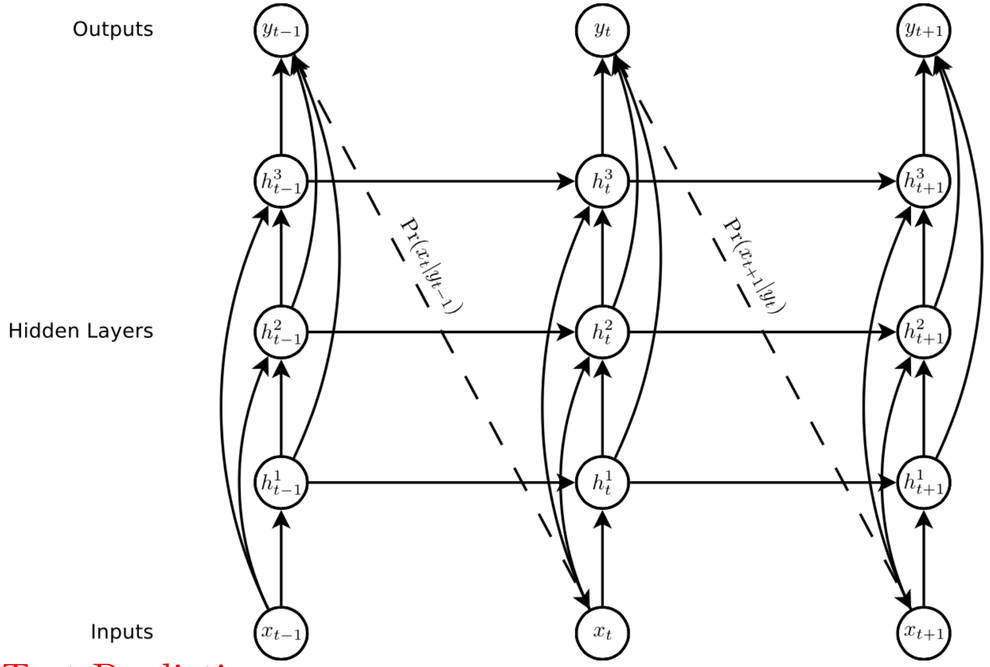
Assistant Professor

Department of Applied Mathematics

University of Colorado Boulder

maziar.raissi@colorado.edu

Generating Sequences With Recurrent Neural Networks



Text Prediction

$x_t \rightarrow$ one-hot-vector (word-level or character-level)

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

$y_t \rightarrow$ output of the network (Stacked RNNs with LSTM cells)

$$\mathcal{L}(\mathbf{x}) = -\sum_{t=1}^{T} \log y_t^{x_{t+1}} \rightarrow \text{loss function}$$

Penn Treebank & Wikipedia Experiments

Evaluation: bits-per-character (BPC) $\rightarrow$ average value of $\log_2 \Pr(x_{t+1}|y_t)$ over the whole test set

perplexity $\rightarrow 2^{5.6\text{BPC}}$ where 5.6 is the average word length (in characters) on the test test

Handwriting Prediction

$x_t \in \mathbb{R}^2 \times \{0, 1\}$

x & y coordinates of the pen offset from the previous inout

end of stroke (yes:1 or no:0)

$$y_t = \left(e_t, \{\pi_t^j, \mu_t^j, \sigma_t^j, \rho_t^j\}_{j=1}^{M}\right) \rightarrow \text{output of the network}$$

$M \rightarrow$ number of mixture components

$e_t \rightarrow$ end of stroke probability

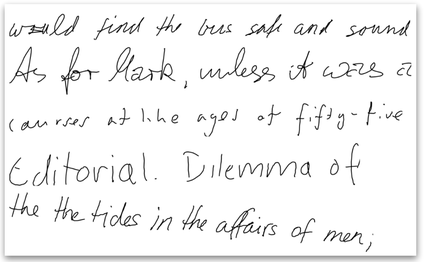
$\pi_t \rightarrow$ mixture weights

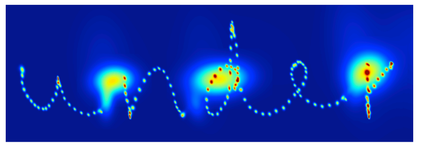
$\mu_t, \sigma_t \rightarrow$ mean and standard deviation (two dimensional)

$\rho_t \rightarrow$ correlation

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

IAM online handwriting database





Synthesis Network

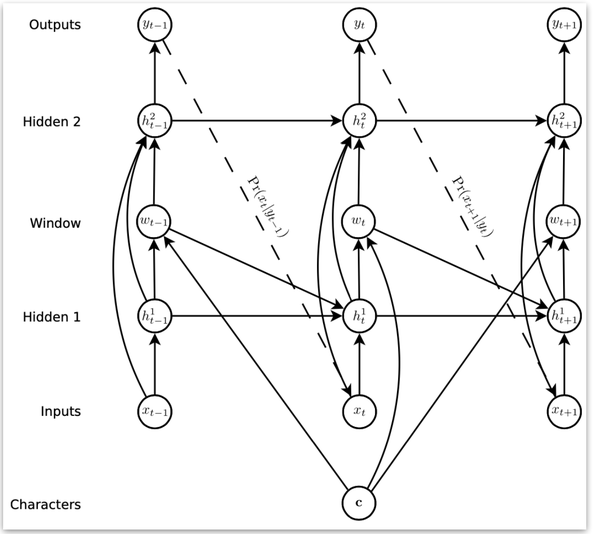
$c \rightarrow$ character sequence of length U

$x \rightarrow$ data sequence of length $T > U$

$$w_t = \sum_{u=1}^{U} \phi(t, u) c_u \rightarrow \text{soft window into } c \text{ at time } t$$

$$\phi(t, u) = \sum_{k=1}^{K} \alpha_t^k \exp\left(-\beta_t^k \left(\kappa_t^k - u\right)^2\right)$$

(window weight of c_u



Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013).

Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling

YouTube Video

– Polyphonic music modeling
– speech signal modeling

RNN

$x = (x_1, x_2, \ldots, x_T)$
$h_0 = 0$
$h_t = \phi(h_{t-1}, x_t), t = 1, \ldots, T$
$y = (y_1, y_2, \ldots, y_T)$
$h_t = g(Wx_t + Uh_{t-1})$

Generative RNN

$p(x_1, \ldots, x_T) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_T \mid x_1, \ldots, x_{T-1})$
$p(x_t \mid x_1, \ldots, x_{t-1}) = g(h_t)$

LSTM

$h_t = o_t \tanh(c_t)$
$c_t \rightarrow$ memory cell
$o_t \rightarrow$ output gate
$o_t = \sigma(W_o x_t + U_o h_{t-1} + V_o c_t)$
$V_o \rightarrow$ diagonal
$c_t = f_t c_{t-1} + i_t \tilde{c}_t$
$f_t \rightarrow$ forget gate
$i_t \rightarrow$ input gate
$\tilde{c}_t \rightarrow$ new memory
$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1})$
$f_t = \sigma(W_f x_t + U_f h_{t-1} + V_f c_{t-1})$
$i_t = \sigma(W_i x_t + U_i h_{t-1} + V_i c_{t-1})$
$V_f, V_i \rightarrow$ diagonal

Gated Recurrent Unit (GRU)

$h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t$
$z_t = \sigma(W_z x_t + U_z h_{t-1}) \rightarrow$ update gate
$\tilde{h}_t = \tanh(Wx_t + U(r_t \odot h_{t-1}))$
$r_t = \sigma(W_r x_t + U_r h_{t-1}) \rightarrow$ reset gate
or $\tilde{h}_t = \tanh(Wx_t + r_t \odot (Uh_{t-1}))$
If $z_t = r_t = 1$ then $h_t = \tilde{h}_t = \tanh(Wx_t + Uh_{t-1})$.
If $z_t = 0$ then $h_t = h_{t-1}$.
If $z_t = 1$ and $r_t = 0$ then $h_t = \tilde{h}_t = \tanh(Wx_t)$.

Sequence modeling

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} \log p\left(x_t^n \mid x_1^n, \ldots, x_{t-1}^n; \boldsymbol{\theta}\right)$$

Polyphonic music modeling datasets

Nottingham, JSB Chorales, MuseData and Piano-midi.

Each symbol in these datasets is respectively a 93-, 96-, 105-, and 108-dimensional binary vector.

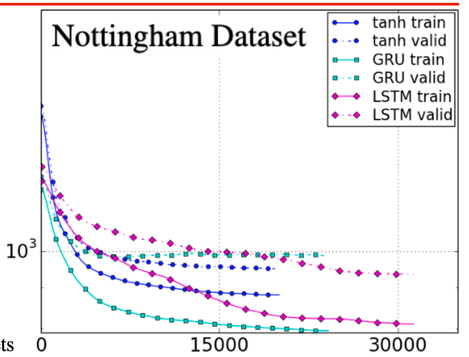
Logistic sigmoid function as output units

Speech signal modeling

Look at 20 consecutive samples to predict the following 10 consecutive samples in a one-dimensional raw audio signal.

Mixture of Gaussians with 20 components as output layer

Unit	# of Units	# of Parameters
Polyphonic music modeling		
LSTM	36	$\approx 19.8 \times 10^3$
GRU	46	$\approx 20.2 \times 10^3$
tanh	100	$\approx 20.1 \times 10^3$
Speech signal modeling		
LSTM	195	$\approx 169.1 \times 10^3$
GRU	227	$\approx 168.9 \times 10^3$
tanh	400	$\approx 168.4 \times 10^3$

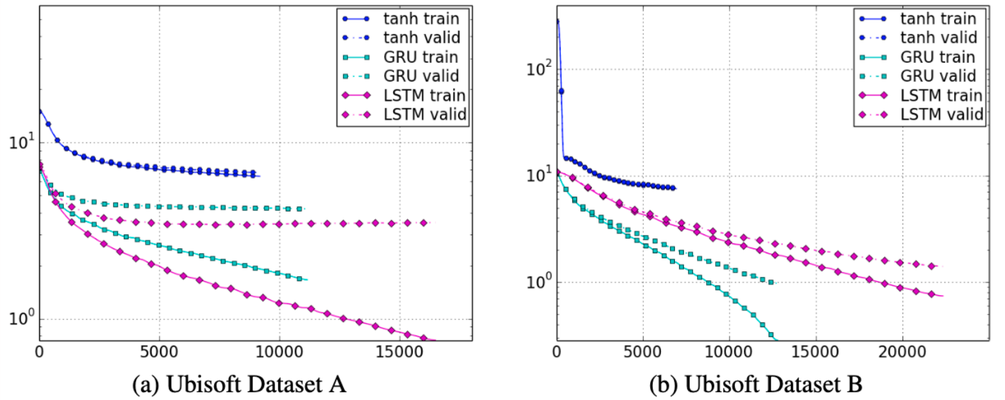


The average negative log-probabilities of the training and test sets

			tanh	GRU	LSTM
Music Datasets	Nottingham	train	3.22	2.79	3.08
		test	**3.13**	3.23	3.20
	JSB Chorales	train	8.82	6.94	8.15
		test	9.10	**8.54**	8.67
	MuseData	train	5.64	5.06	5.18
		test	6.23	**5.99**	6.23
	Piano-midi	train	5.64	4.93	6.49
		test	9.03	**8.82**	9.03
Ubisoft Datasets	Ubisoft dataset A	train	6.29	2.31	1.44
		test	6.44	3.59	**2.70**
	Ubisoft dataset B	train	7.61	0.38	0.80
		test	7.62	**0.88**	1.26

Results are not conclusive in comparing LSTM and GRU!

Wall Clock Time (seconds)



(a) Ubisoft Dataset A (b) Ubisoft Dataset B

Chung, Junyoung, et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling." *arXiv preprint arXiv:1412.3555* (2014).

WaveNet: A Generative Model for Raw Audio

Boulder

YouTube Playlist

$x = \{x_1, x_2, \ldots, x_T\} \rightarrow$ waveform

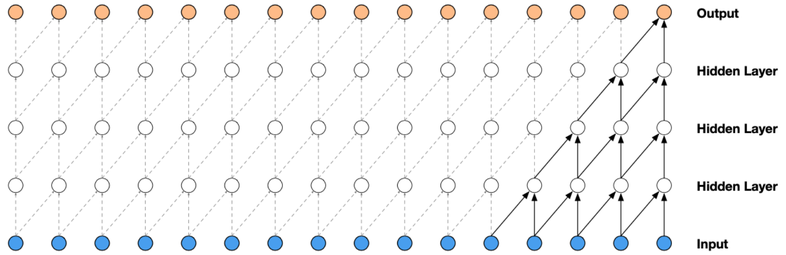
$$p(x) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1})$$

$p(x_t | x_1, \ldots, x_{t-1}) \rightarrow$ modeled by a stack of convolutional layers

$x_t \rightarrow$ audio sample

categorical distribution over the next value x_t with a softmax layer

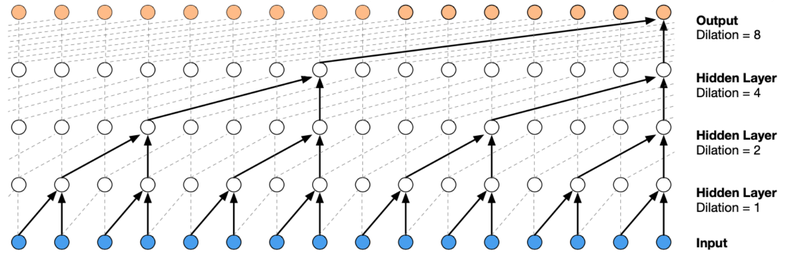
Causal Convolutions



receptive field $= 5 = (\underbrace{\#\text{ layers}}_{=4} + \underbrace{\text{filter length}}_{=2} - 1)$

Dilated Causal Convolutions

à trous convolutions (convolutions with holes)



$\underbrace{1, 2, 4, \ldots, 512}_{\text{receptive field size } 1024}, 1, 2, 4, \ldots, 512, 1, 2, 4, \ldots, 512 \rightarrow$ dilations

Softmax Distributions

sequence of 16-bit integer values (one per time step)

$2^{16} = 65,536$ probabilities per time step $\rightarrow$ softmax

$$f(x_t) = \text{sign}(x_t) \frac{\ln(1 + \mu|x_t|)}{1 + \mu} \rightarrow \mu\text{-law companding transformation}$$

$-1 < x_t < 1$ and $\mu = 255$

Quantize the transformed output to 256 possible values

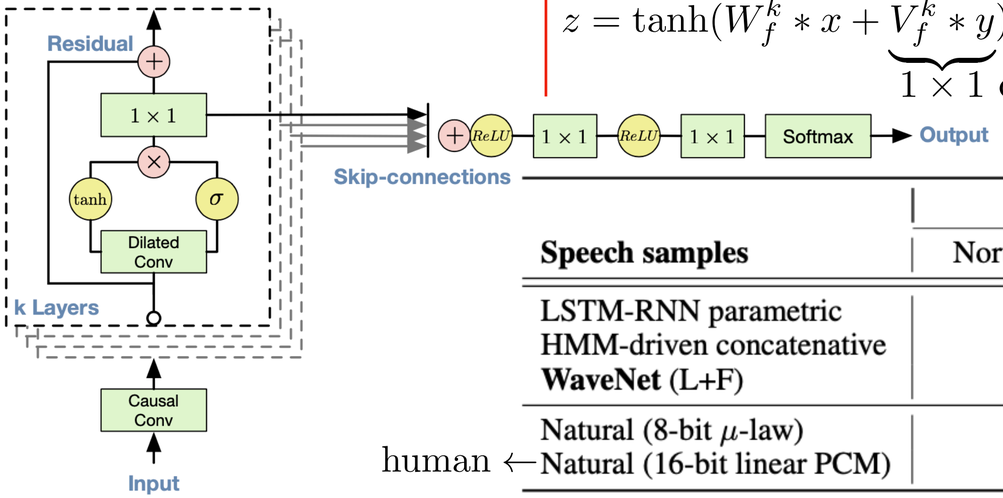
Gated Activation Units

$$z = \tanh(W_f^k * x) \odot \sigma(W_g^k * x)$$

$k \rightarrow$ layer index

f & $g \rightarrow$ filter & gate

Residual and Skip Connections



Conditional WaveNets

$h \rightarrow$ additional input

multi-speaker setting: speaker identity

text-to-speech (TTS): text

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_1, \ldots, x_{t-1}, h)$$

Global Conditioning

$$z = \tanh(W_f^k * x + \underbrace{(V_f^k)^T h}_{\text{broadcast over all time dimensions}}) \odot \sigma(W_g^k * x + (V_g^k)^T h)$$

Local Conditioning

$h = \{h_t\} \rightarrow$ lower sampling frequency than the audio signal

$y = f(h)$

$y \rightarrow$ same resolution as the audio signal

$f \rightarrow$ learned upsampling (transposed convolutional network)

$$z = \tanh(W_f^k * x + \underbrace{V_f^k * y}_{1 \times 1 \text{ convolution}}) \odot \sigma(W_g^k * x + V_g^k * y)$$

text-to-speech

Speech samples	Subjective 5-scale MOS in naturalness: North American English	Subjective 5-scale MOS in naturalness: Mandarin Chinese
LSTM-RNN parametric	3.67 ± 0.098	3.79 ± 0.084
HMM-driven concatenative	3.86 ± 0.137	3.47 ± 0.108
WaveNet (L+F)	$\mathbf{4.21} \pm 0.081$	$\mathbf{4.08} \pm 0.085$
Natural (8-bit μ-law)	4.46 ± 0.067	4.25 ± 0.082
human $\leftarrow$ Natural (16-bit linear PCM)	4.55 ± 0.075	4.21 ± 0.071

Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).

Boulder

Questions?

YouTube Playlist