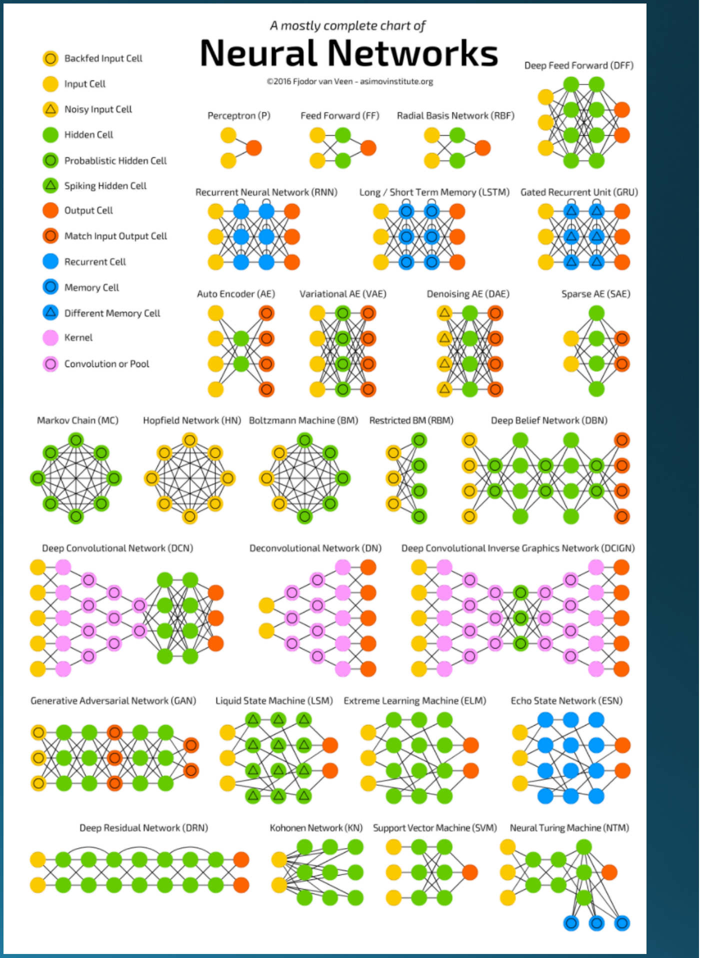
CSCI 4360/6360 Data Science II

Recurrent Neural Networks

The Neural Network Zoo

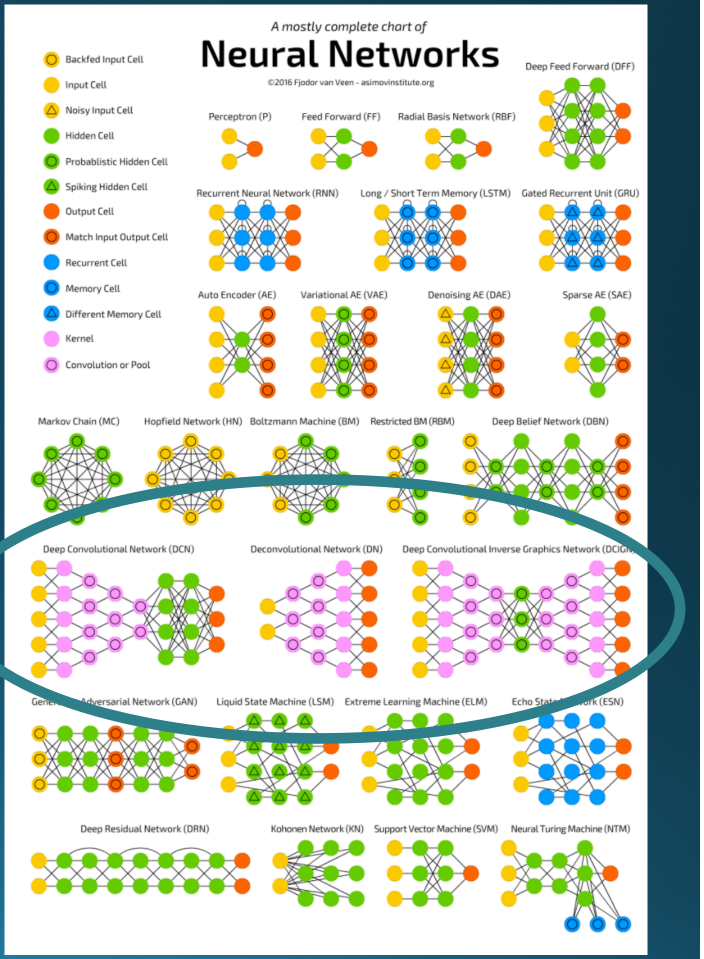
- http://www.asimovinstitute.org/neural-network-zoo/


A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

The Neural Network Zoo

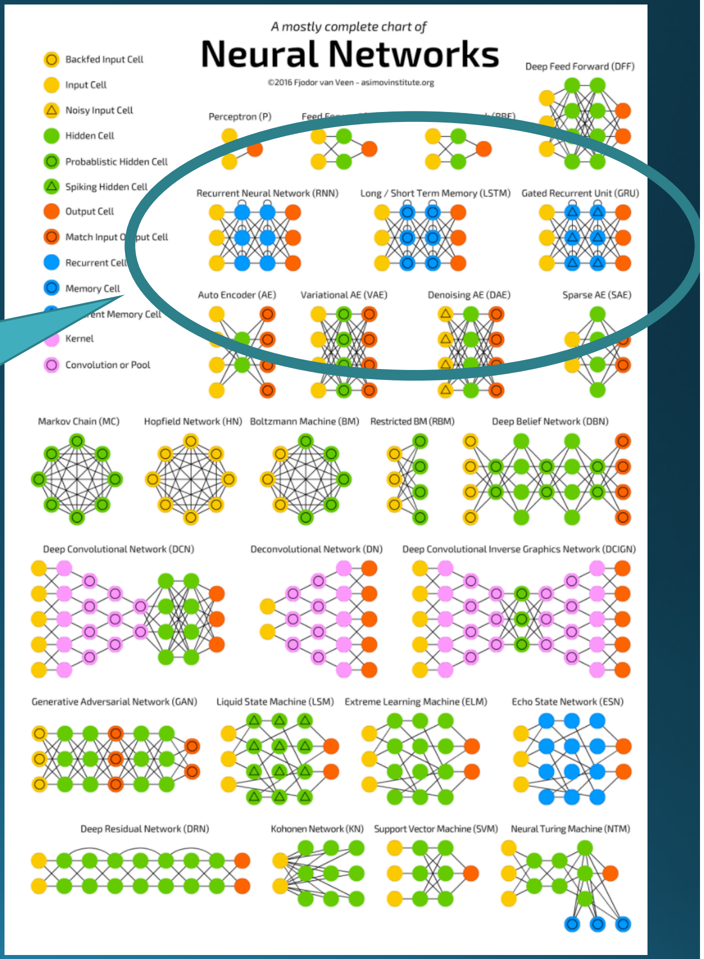
- http://www.asimovinstitute.org/neural-network-zoo/

Last time


A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
Different Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Forward (FF)
Radial Basis Network (RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN
Gener
Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo Stat
ork (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

The Neural Network Zoo

- http://www.asimovinstitute.org/neural-network-zoo/


A mostly complete chart of
Neural Networks
©2016 Fjodor van Veen - asimovinstitute.org
Backfed Input Cell
Input Cell
Noisy Input Cell
Hidden Cell
Probablistic Hidden Cell
Spiking Hidden Cell
Output Cell
Match Input Output Cell
Recurrent Cell
Memory Cell
ent Memory Cell
Kernel
Convolution or Pool
Perceptron (P)
Feed Fo
(RBF)
Deep Feed Forward (DFF)
Recurrent Neural Network (RNN)
Long / Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Auto Encoder (AE)
Variational AE (VAE)
Denoising AE (DAE)
Sparse AE (SAE)
Markov Chain (MC)
Hopfield Network (HN)
Boltzmann Machine (BM)
Restricted BM (RBM)
Deep Belief Network (DBN)
Deep Convolutional Network (DCN)
Deconvolutional Network (DN)
Deep Convolutional Inverse Graphics Network (DCIGN)
Generative Adversarial Network (GAN)
Liquid State Machine (LSM)
Extreme Learning Machine (ELM)
Echo State Network (ESN)
Deep Residual Network (DRN)
Kohonen Network (KN)
Support Vector Machine (SVM)
Neural Turing Machine (NTM)

Today

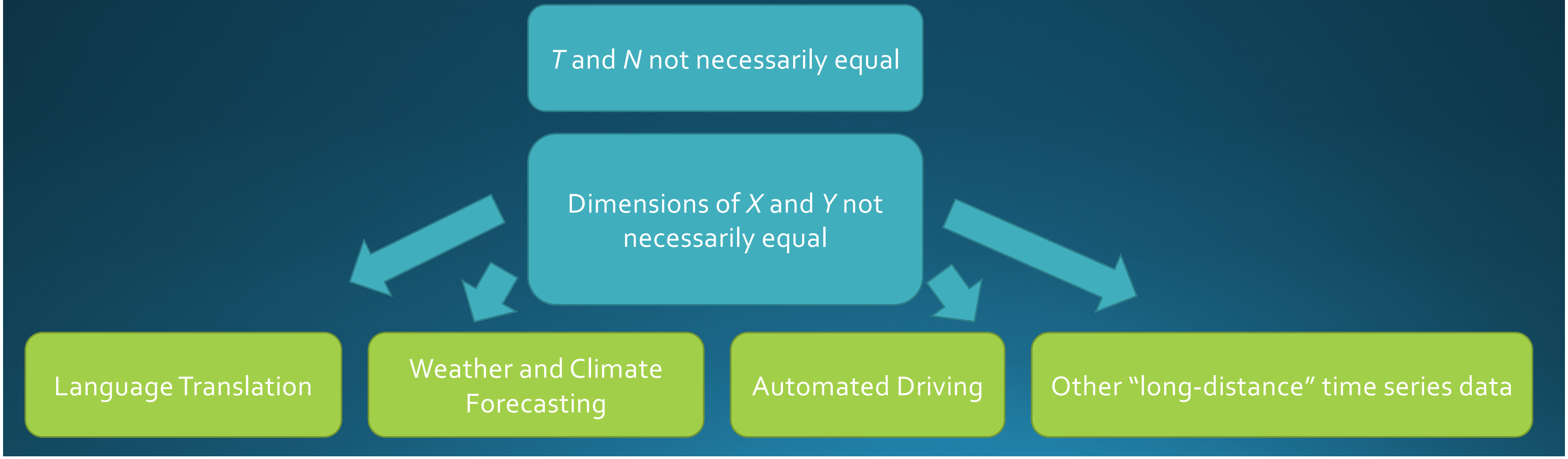
Modeling Sequences

- Input:

$$X = [\vec{x}_1, \vec{x}_2, ..., \vec{x}_T]$$

- Output:

$$Y = [\vec{y}_1, \vec{y}_2, ..., \vec{y}_N]$$



T and *N* not necessarily equal

Dimensions of *X* and *Y* not necessarily equal

Language Translation

Weather and Climate Forecasting

Automated Driving

Other “long-distance” time series data

Something we’ve seen before

Linear Dynamical Models

- Two main components (using notation from Hyndman 2006):

Appearance Model

$$y_t = Cx_t + u_t$$

State Model

$$x_t = Ax_{t-1} + Wv_t$$

Autoregressive Models

- This is the definition of a 1st-order autoregressive (AR) process!

$$x_t = Ax_{t-1} + Wv_t$$

- Each observation (x_t) is a function of previous observations, plus some noise

- **Markov model!**

Autoregressive Models

- AR models can have higher orders than 1
- Each observation is dependent on the previous *d* observations

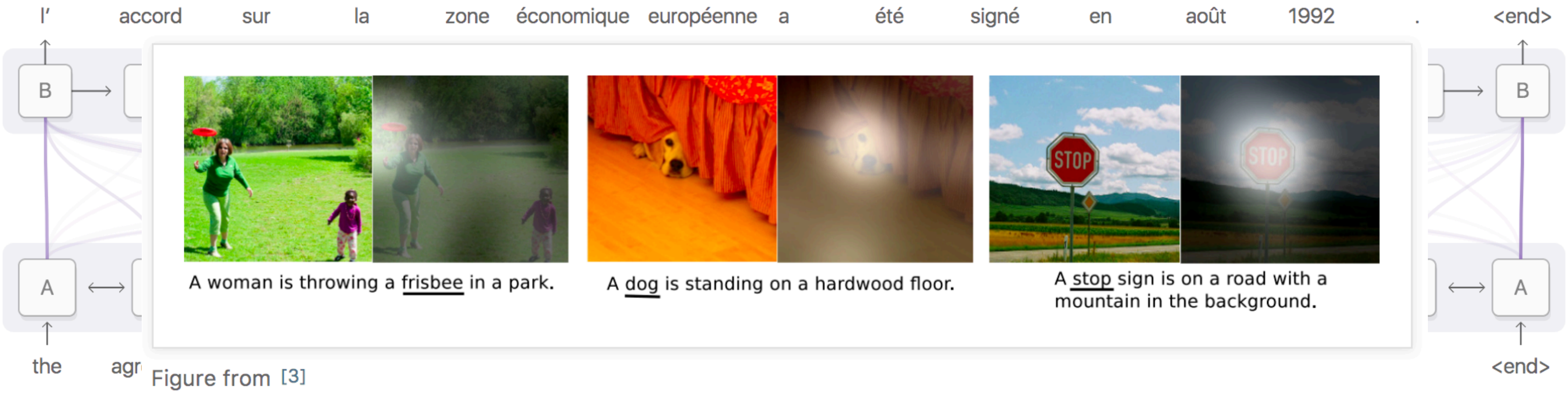
$$x_t = A_1 x_{t-1} + A_2 x_{t-2} + \dots + A_d x_{t-d} + W v_t$$

Autoregressive Models

- Concrete, *a priori* definition of what is important
  - nth-order Markov process
  - n+1 terms and larger are explicitly ignored
- No concept of *attention*
  - All *n* terms receive equal "attention" (computationally, if not also statistically)
  - Are you devoting equal time reading every word on this slide?
- Cannot handle *variable-length inputs*, nor *variable-length outputs*
  - Contrast with CNNs: all input images have to be the same size (usually)
  - Contrast with [insert deep network of choice]: all outputs are the same, given any input

Attention

- Some things are more important than others



A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

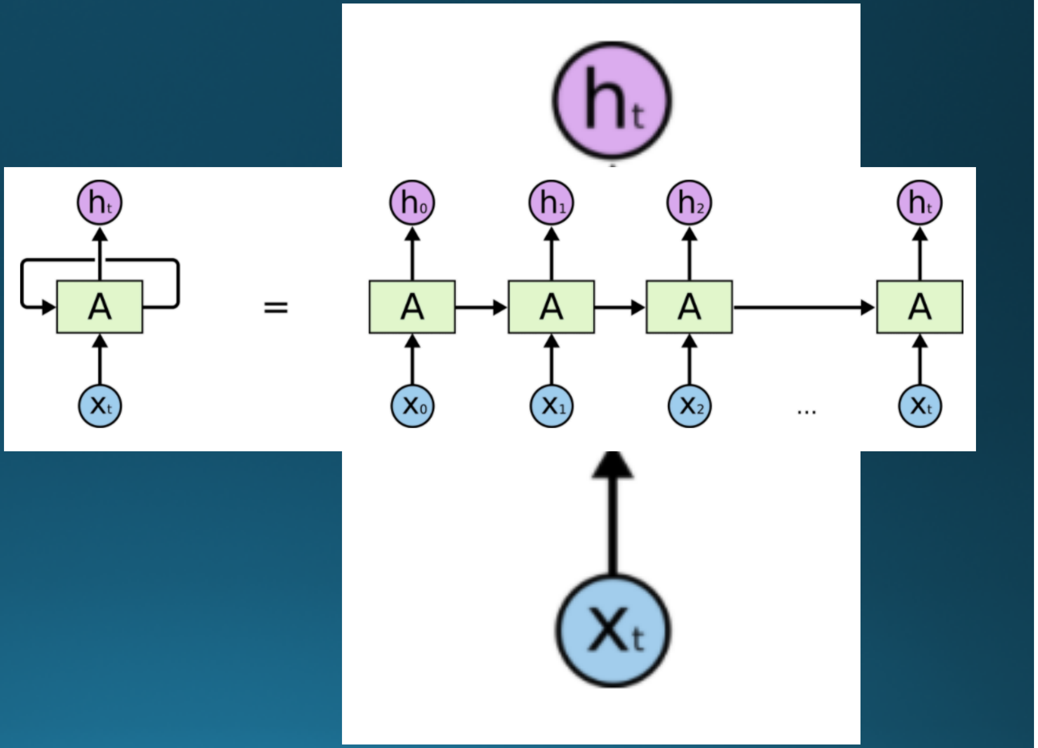
A stop sign is on a road with a mountain in the background.

Figure from [3]

Diagram derived from Fig. 3 of Bahdanau, *et al.* 2014

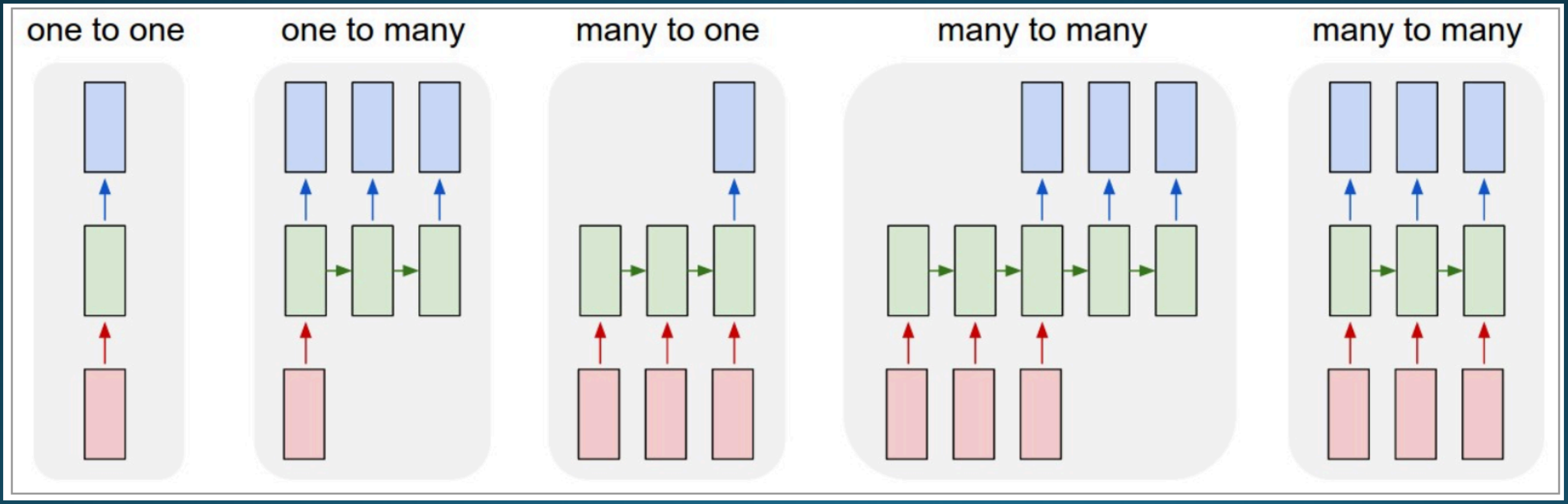
Recurrent Neural Networks

- In short, recurrent neural networks (RNNs) break the typical "directed acyclic" pedagogy of deep networks by introducing self-loops
  - Allows information to persist through multiple iterations
- We can get around problems introduced by loops by "unrolling" the loops
  - This permits backprop to work as usual



Recurrent Neural Network

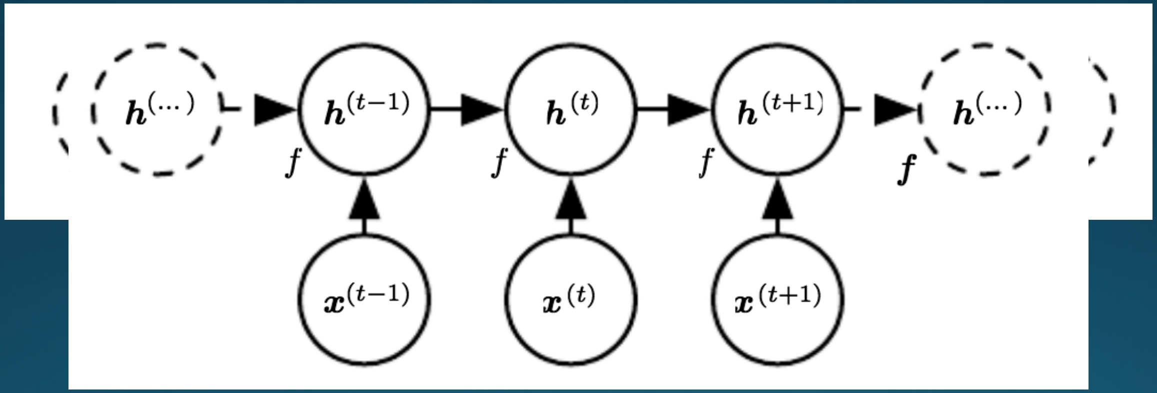
- “List” structure intrinsically handles variable-length data



- Think: convolution, but over time instead of space

Recurrent Neural Networks

- Use the same "parameter sharing" as CNNs
  - And linear dynamical systems!



- *f* maps each time point to the next
- Also updates internal state *h*

Recurrent Neural Networks

- Four main equations at each time point

$$\vec{a}^{(t)} = \vec{b} + W\vec{h}^{(t-1)} + U\vec{x}^{(t)}$$

$$\vec{h}^{(t)} = \sigma(\vec{a}^{(t)})$$

$$\vec{o}^{(t)} = \vec{c} + V\vec{h}^{(t)}$$

$$\hat{\vec{y}}^{(t)} = \phi(\vec{o}^{(t)})$$

Bias term

Internal RNN state

Input

Weights for hidden-to-hidden connections

Update internal state

Activation function

Bias term

Internal RNN state

Predicted output

Final layer activation

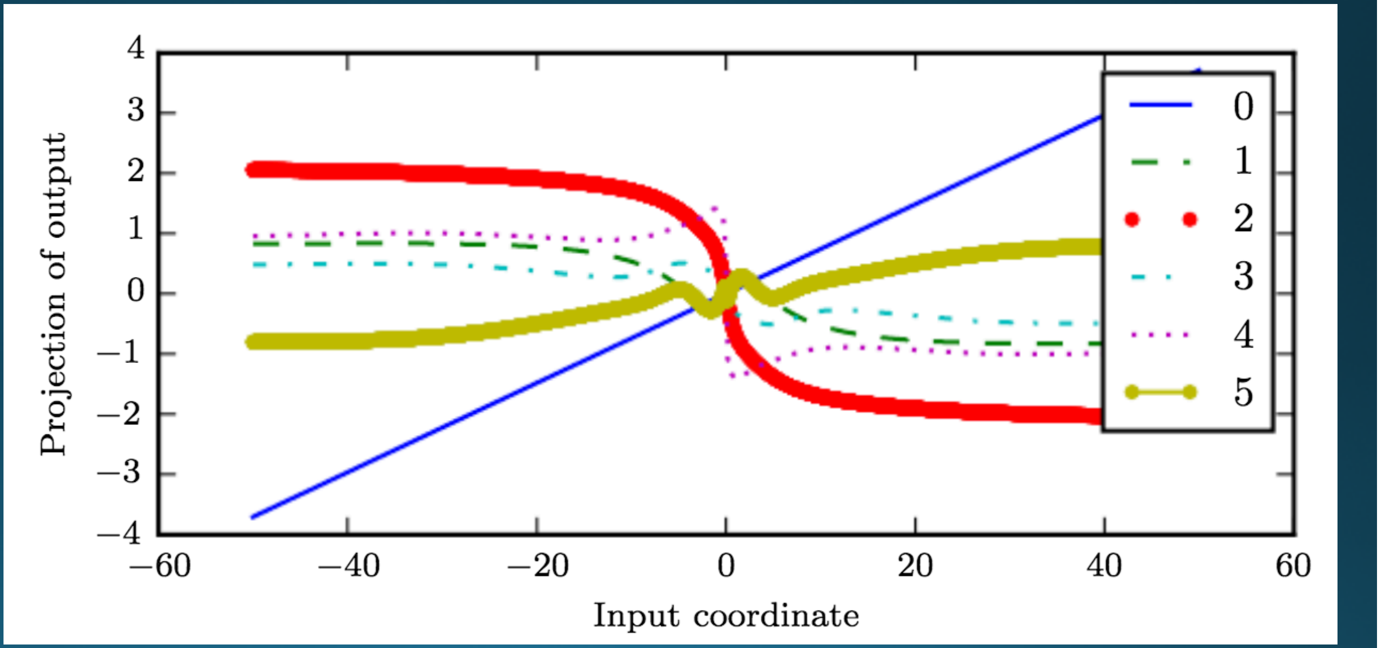
hidden-to-output connections

Recurrent Neural Networks

- RNNs are great for modeling sequences, but by themselves cannot capture *attention*
- **Long-term dependencies require an explicit “memory”**

Long-term Dependencies

- RNNs *compose* the same activation function repeatedly
  - Think: recurrence relations
- Results in highly nonlinear behavior



Long-term Dependencies

- Put another way, recall the internal state update:

$$\vec{h}^{(t)} = W^T \vec{h}^{(t-1)}$$

- Where have we seen this before…

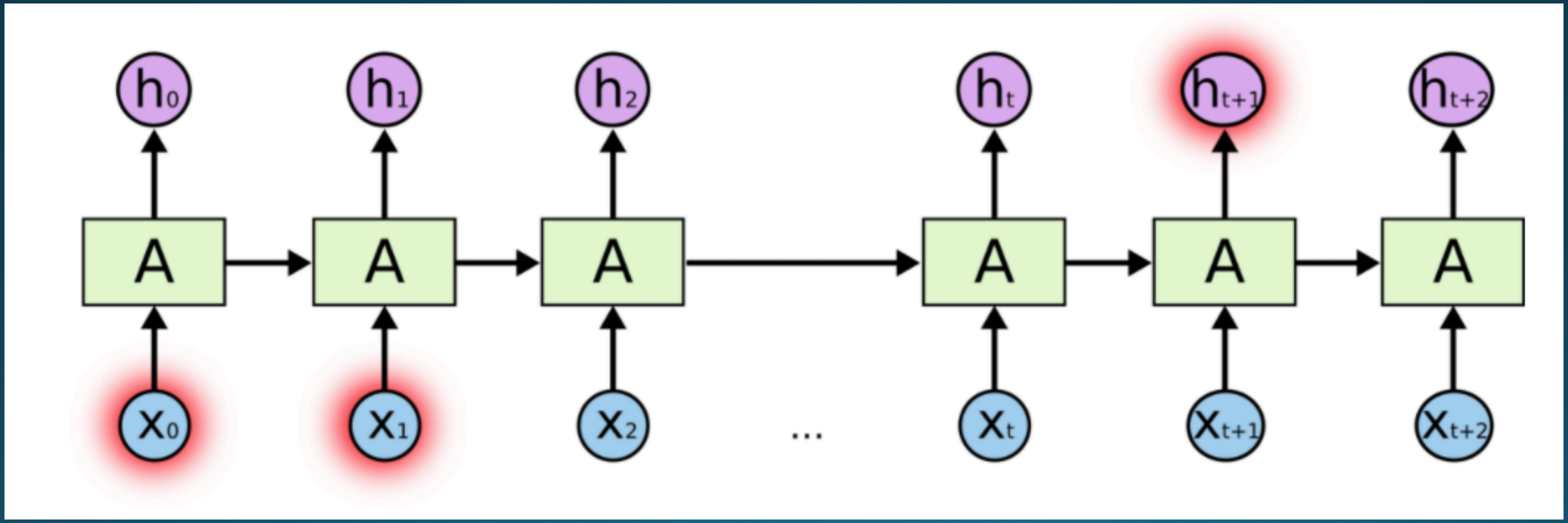
$$\vec{h}^{(t)} = (W^t)^T \vec{h}^{(0)} \qquad W = X \Lambda X^T$$

$$\boxed{\vec{h}^{(t)} = X^T \Lambda^t X \vec{h}^{(0)}}$$

- Eigenvalues are raised to the power *t*, decaying any eigenvalue < 1
- **Any component of *h*$^{(0)}$ not aligned with largest eigenvalue will be discarded**

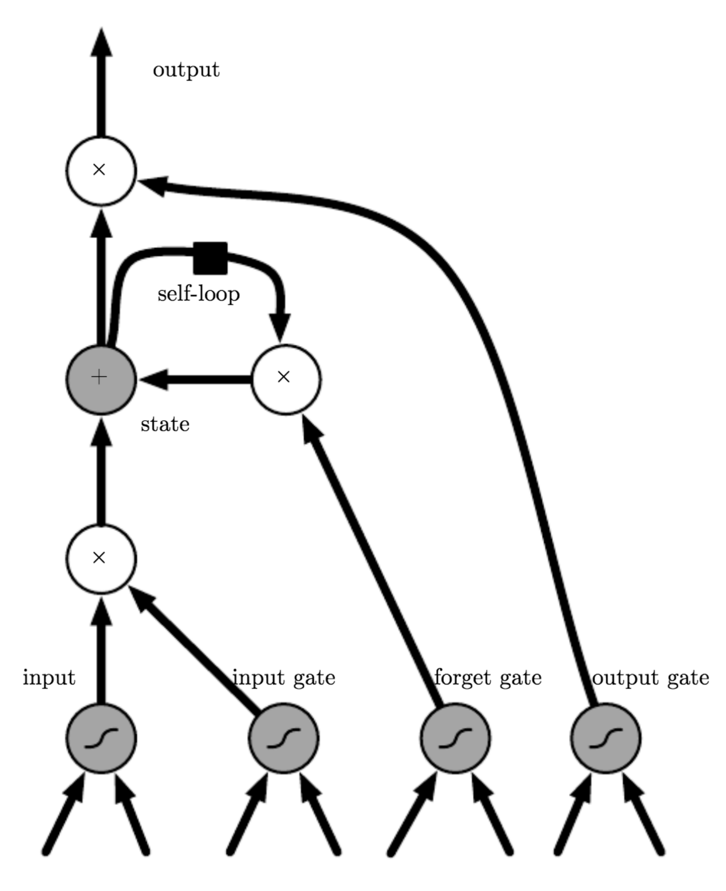
Long-term Dependencies

- “I grew up in France… I speak fluent **French**.”



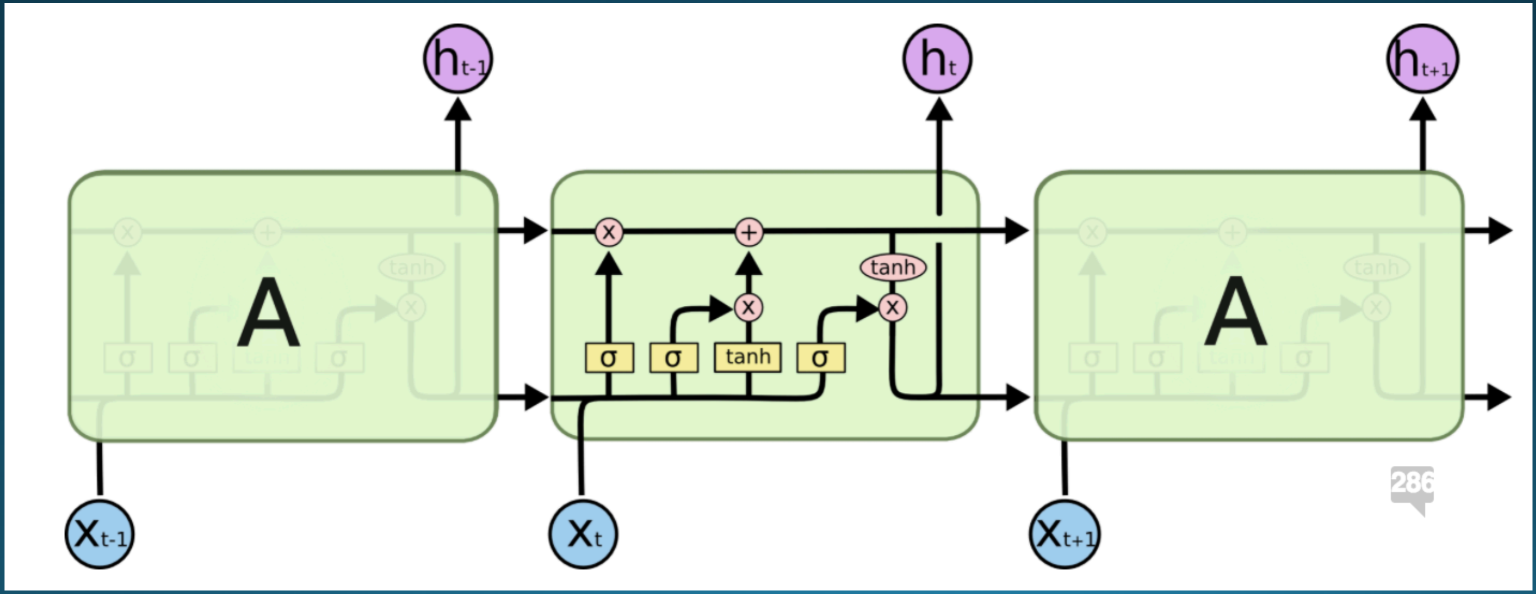
Long-Short Term Memory

- Or “LSTM”
- A variant of the *gated* RNN
- Each hidden state comprises a **forget** gate
  - Determines what to “remember” and what to discard
  - Functions on self-loop input



LSTM versus “vanilla” RNN

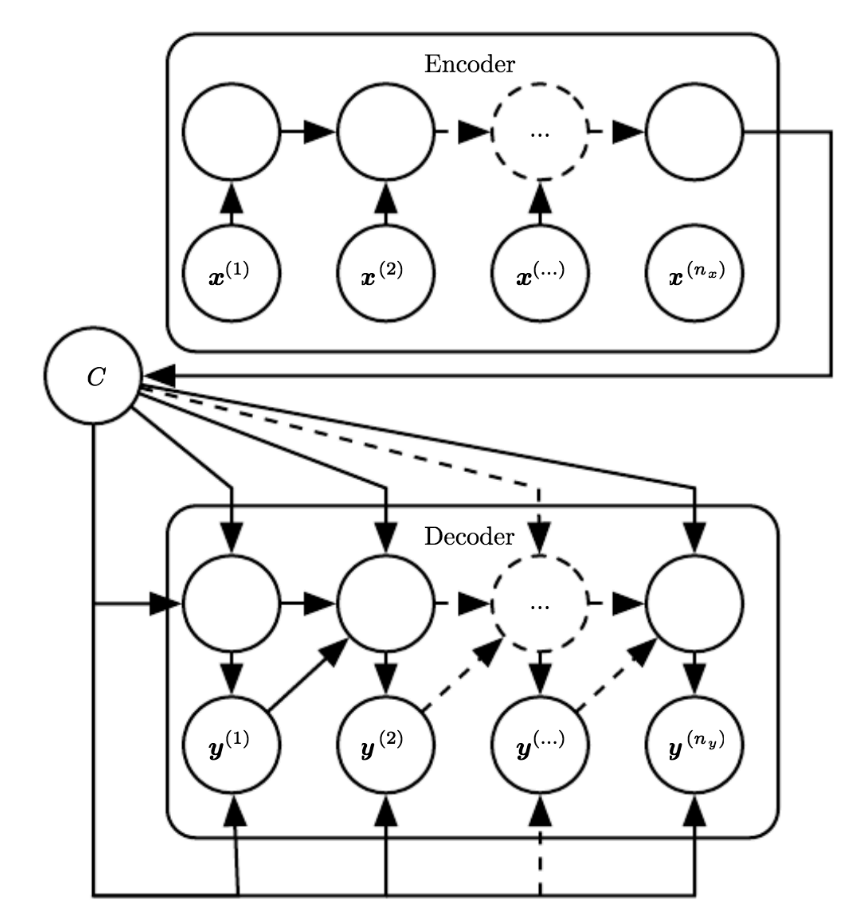
- A “vanilla” RNN contains only a single activation
- LSTMs have four interacting layers in each step



Other RNN Variants

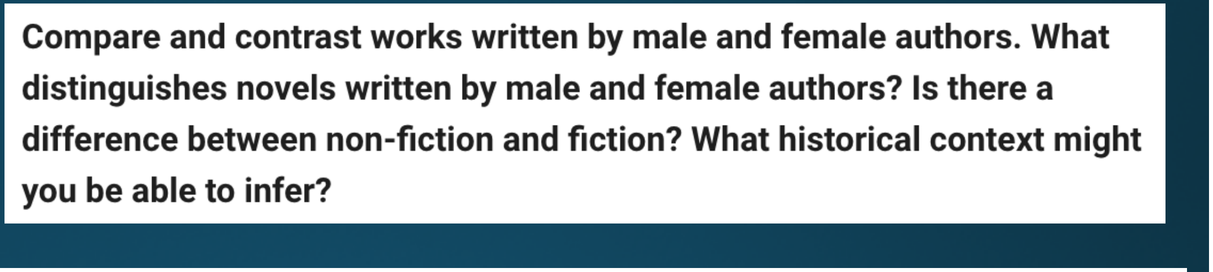
Encoder-Decoder Networks

- Maps input to output sequences
  - Each mapping not necessarily of equal length!
- *C* is a "semantic summary"
  - Think: input "subspace"
- Have to ensure *C* is of sufficient dimensionality to represent input space

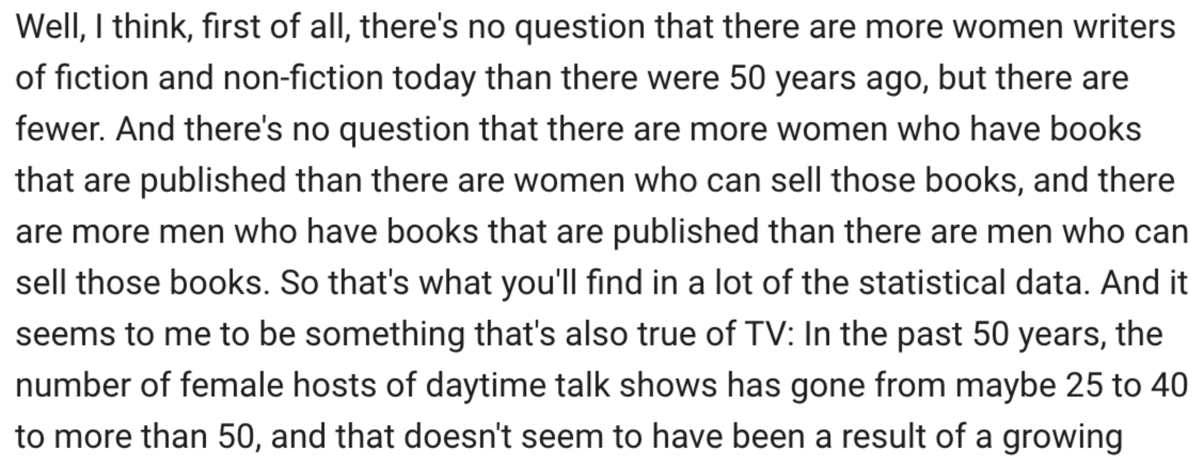


Transformer Networks

- Architectural implementation of “self-attention”
  - GPT-2, XLM, MultiFit, BERT
- “Talk to Transformer” https://talktotransformer.com/
- **Language models**



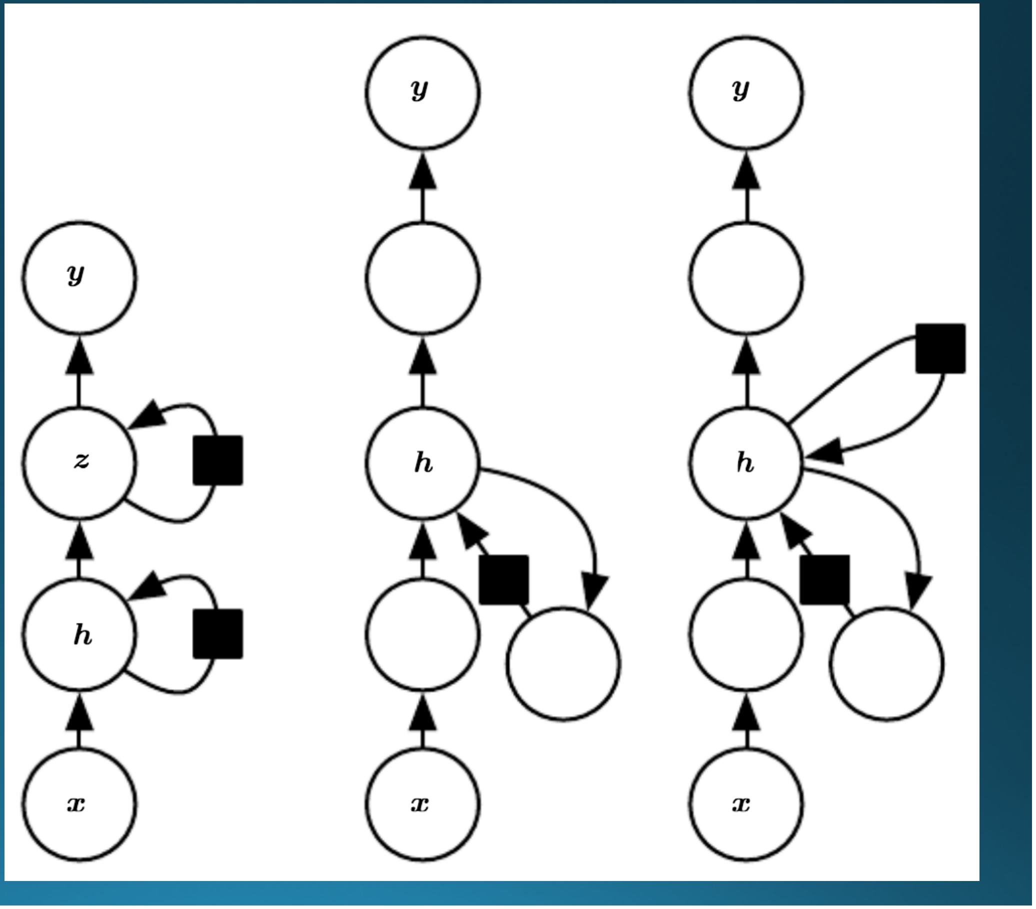
Compare and contrast works written by male and female authors. What distinguishes novels written by male and female authors? Is there a difference between non-fiction and fiction? What historical context might you be able to infer?



Well, I think, first of all, there's no question that there are more women writers of fiction and non-fiction today than there were 50 years ago, but there are fewer. And there's no question that there are more women who have books that are published than there are women who can sell those books, and there are more men who have books that are published than there are men who can sell those books. So that's what you'll find in a lot of the statistical data. And it seems to me to be something that's also true of TV: In the past 50 years, the number of female hosts of daytime talk shows has gone from maybe 25 to 40 to more than 50, and that doesn't seem to have been a result of a growing

Deep Recurrent Networks

- Each recurrent state can feed into a series of hidden states
- Analogous to hidden markov models (HMMs) with attention and nearly infinite support for hidden states

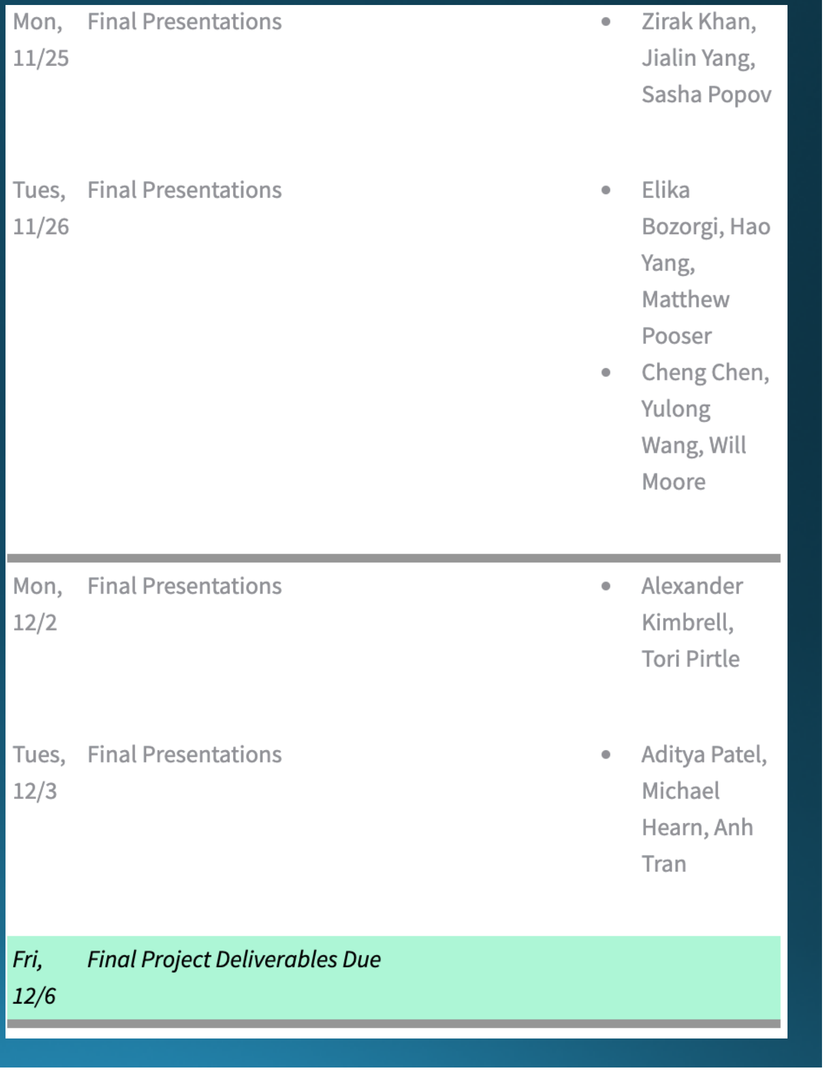


Conclusions

- Recurrent neural networks
  - A generalization of convolution (or is a convolution a generalization of recurrence?): uses same **parameter-sharing** idea
  - Introduces self-loops, but over discrete intervals: loops can be "unrolled" so backpropagation can still be used as normal
  - Still have trouble with long-term dependencies, such as language translation (vanishing / exploding gradient)
- Long-short term memory
  - Introduce a series of gates within the self-loops
  - Gates determine what to remember, what to discard
  - No ill-conditioned gradients
- Other gated variants

Course Details

- Assignment 5—due **tonight!**
- If you're getting ZEROs on AutoLab:
  - Check your **dimensions**
  - (p and q have NO EFFECT on dimensions, only k)
  - A/B/C RMSD are scaled to expected
    - > 100: you somehow made something more accurate than the baseline?! (check that you are correctly using random numbers)
    - < 100: Double-check linalg operations (and make sure you're using **scipy.linalg**)
- Final Projects
  - Current presentation schedule
  - 25 minutes (+5 for Q&A) per team



Mon, 11/25	Final Presentations	• Zirak Khan, Jialin Yang, Sasha Popov
Tues, 11/26	Final Presentations	• Elika Bozorgi, Hao Yang, Matthew Pooser • Cheng Chen, Yulong Wang, Will Moore
Mon, 12/2	Final Presentations	• Alexander Kimbrell, Tori Pirtle
Tues, 12/3	Final Presentations	• Aditya Patel, Michael Hearn, Anh Tran
Fri, 12/6	*Final Project Deliverables Due*	

References

- Deep Learning Book, Chapter 10: "Sequence Modeling: Recurrent and Recursive Nets", http://www.deeplearningbook.org/contents/rnn.html
- "Attention and Augmented Recurrent Neural Networks", https://distill.pub/2016/augmented-rnns/
- "Understanding LSTM Networks" https://colah.github.io/posts/2015-08-Understanding-LSTMs/
- "The Unreasonable Effectiveness of Recurrent Neural Networks" https://karpathy.github.io/2015/05/21/rnn-effectiveness/
- "MultiFiT: Efficient Multi-lingual Language Model Fine-tuning", https://arxiv.org/abs/1909.04761
- "Cross-lingual Language Model Pretraining", https://arxiv.org/abs/1901.07291
- "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", https://arxiv.org/abs/1810.04805
- Full GPT-2 Model https://openai.com/blog/gpt-2-1-5b-release/