

Sequence Modelling and Recurrent Neural Networks (RNNs)

Pattern Recognition

Fall 2019

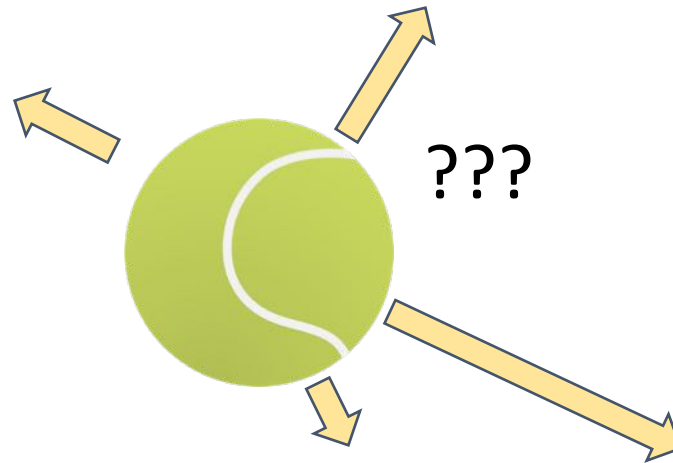
Dennis Madsen

Topic overview

- Neural Networks (recap) and Deep Learning
- Improving DNN: Hyperparameter tuning, regularization, optimization
- Convolutional Neural Networks (CNN)
- CNN popular architectures
- **Sequence Models/Recurrent neural networks (RNN)**
 - RNN architecture
 - Gated cells for long term dependencies
 - Natural Language Processing (NLP)
- **Beyond the basics (object detection and segmentation)**
 - Neural networks as a generative model

Motivation

Given a single image instance of a ball - can we predict its direction?



Motivation

Given a single image instance of a ball - can we predict its direction?



Given enough previous placements, it is possible to predict direction and magnitude

Examples of sequence data

Speech recognition



→ "The quick brown fox jumped over the lazy dog."

Music generation

∅



Sentiment classification

"There is nothing to like in this movie."



Machine translation

Vil du synge med mig?

→ Do you want to sing with me?

Video activity recognition



→ Running

Name entity recognition

Yesterday, Harry Potter met Hermione Granger.

→ Yesterday, **Harry Potter** met **Hermione Granger**.

All of these problems can be addressed using supervised learning with labeled data

Words: One-hot representation

Words are represented as a one-hot feature representation.

The vocabulary is therefore fixed with words representing a single entry in a vector.

For commercial applications, vocabularies of 30-50.000 words are often used.

x: "Last week I visited Paris, the capital of France."

$x^{<1>}$

$x^{<5>}$

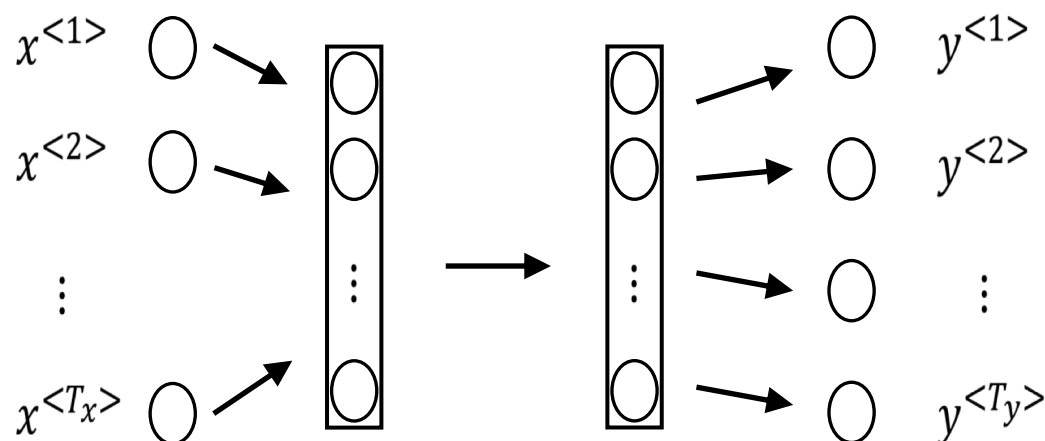
$x^{<10>}$

Rome	=	[1, 0, 0, 0, 0, 0, ..., 0]
Paris	=	[0, 1, 0, 0, 0, 0, ..., 0]
Italy	=	[0, 0, 1, 0, 0, 0, ..., 0]
France	=	[0, 0, 0, 1, 0, 0, ..., 0]

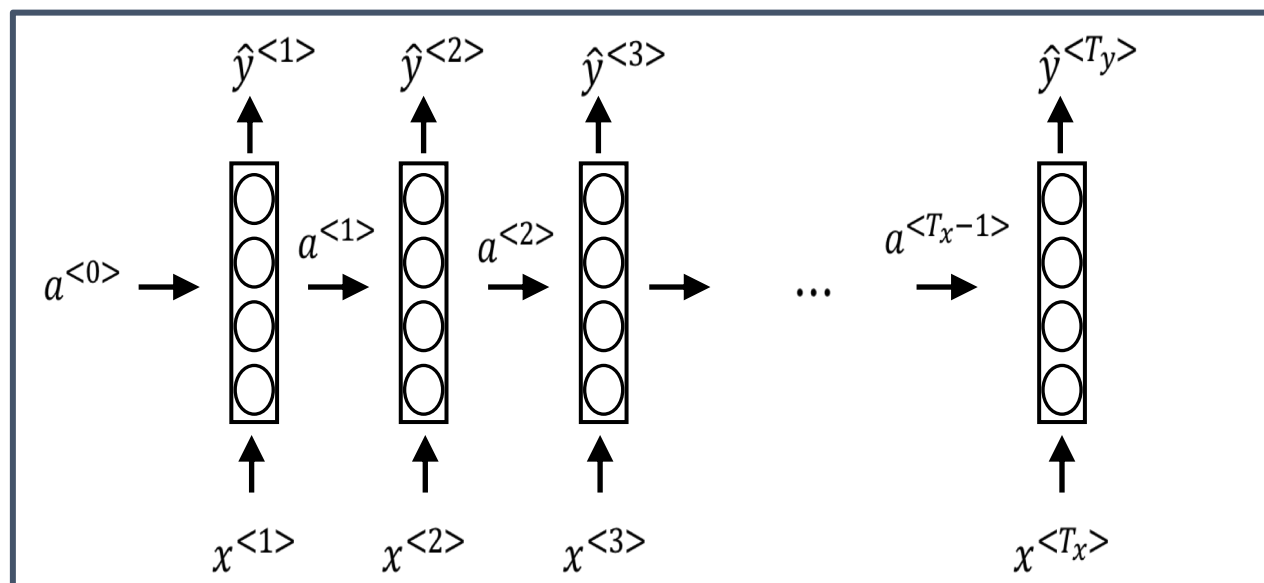
Why not a standard fully connected NN?

Problems:

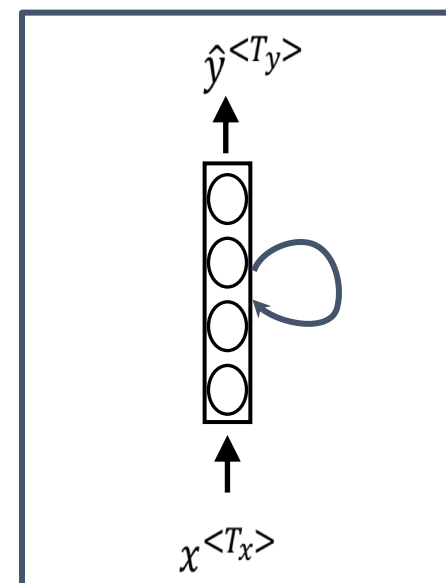
- Input/outputs can be different lengths in different examples.
 - Example: language translation doesn't happen word to word.
- Does not share features across different locations (bag-of-words cannot be used).
 - The food was good, not bad at all (positive).
 - The food was bad, not good at all (negative).



Recurrent Neural network structure



Unrolled representation



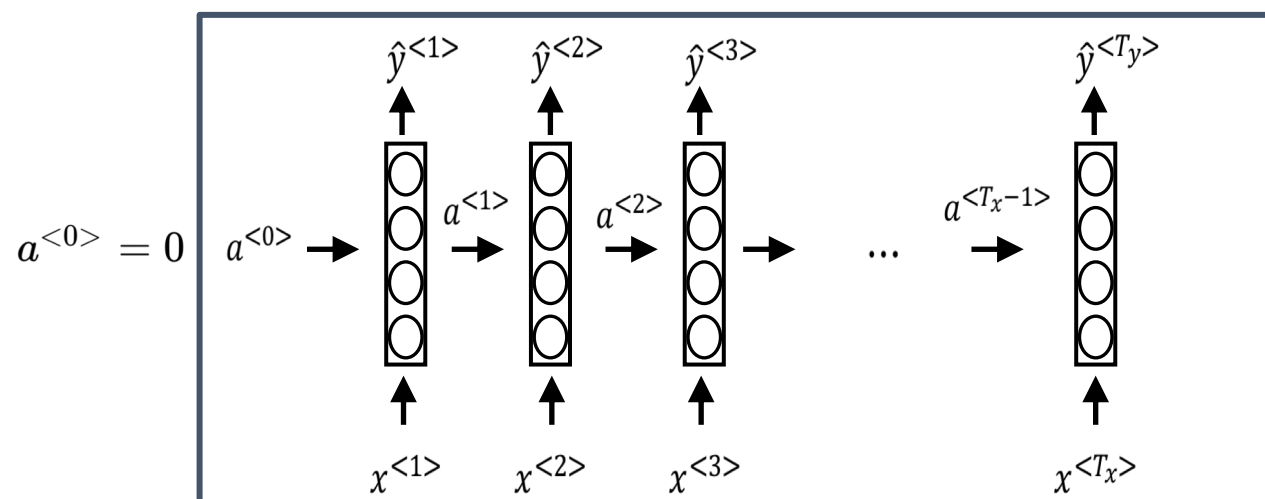
Compact representation

RNN notation

- $g()$ is the activation function such as: tanh, sigmoid, ReLU.

$$a^{<1>} = g(W_{aa}a^{<0>} + W_{ax}x^{<1>} + b_a)$$

$$\hat{y}^{<1>} = g(W_{ya}a^{<1>} + b_y)$$



$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

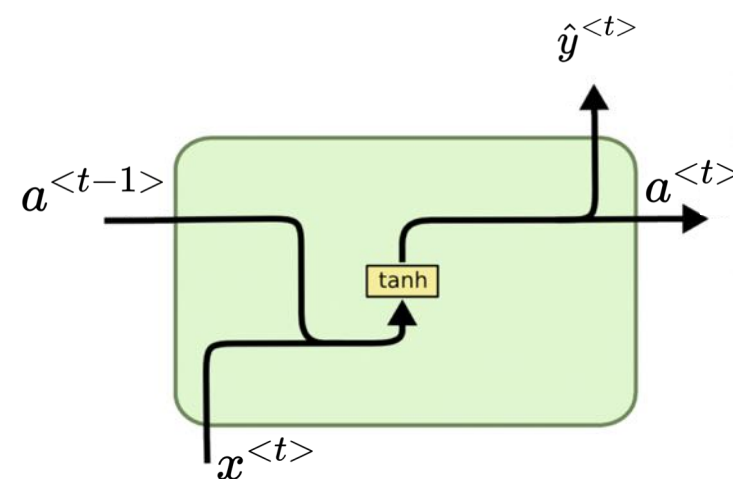
$$W_a = [W_{aa} | W_{ax}]$$

$$\hat{y}^{<t>} = g(W_{ya}a^{<t>} + b_y)$$

- First subscript of w defines the output of the multiplication.
- Second subscript of w defines what it is being multiplied with.

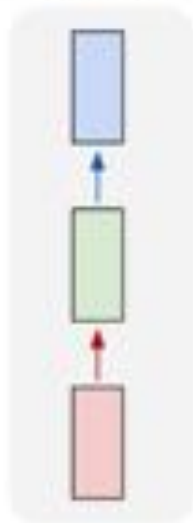
Types of RNNs

- One to One: "Vanilla" neural network.
- One to Many: Music generation.
- Many to One: Sentiment classification.
- Many to Many: Translation.

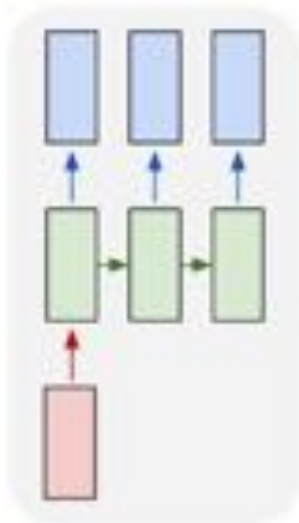


$$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$$

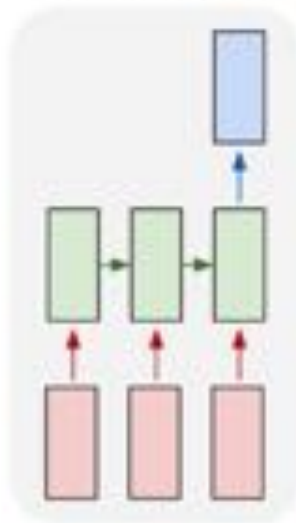
one to one



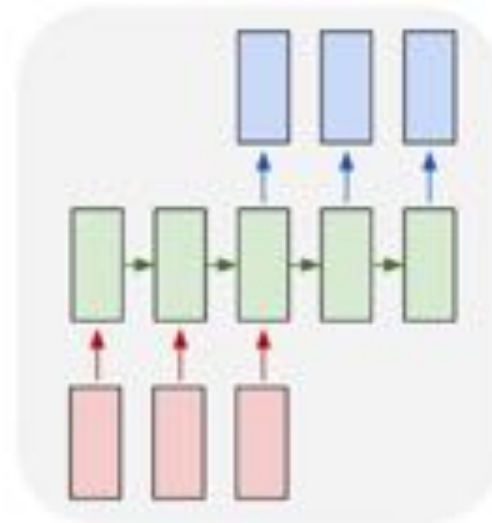
one to many



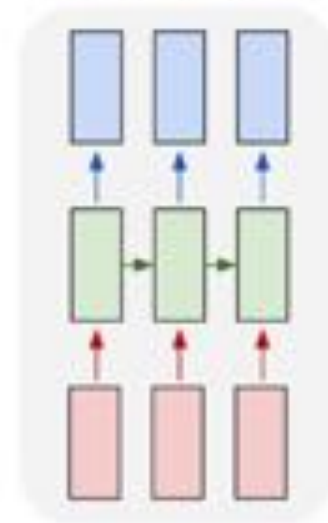
many to one



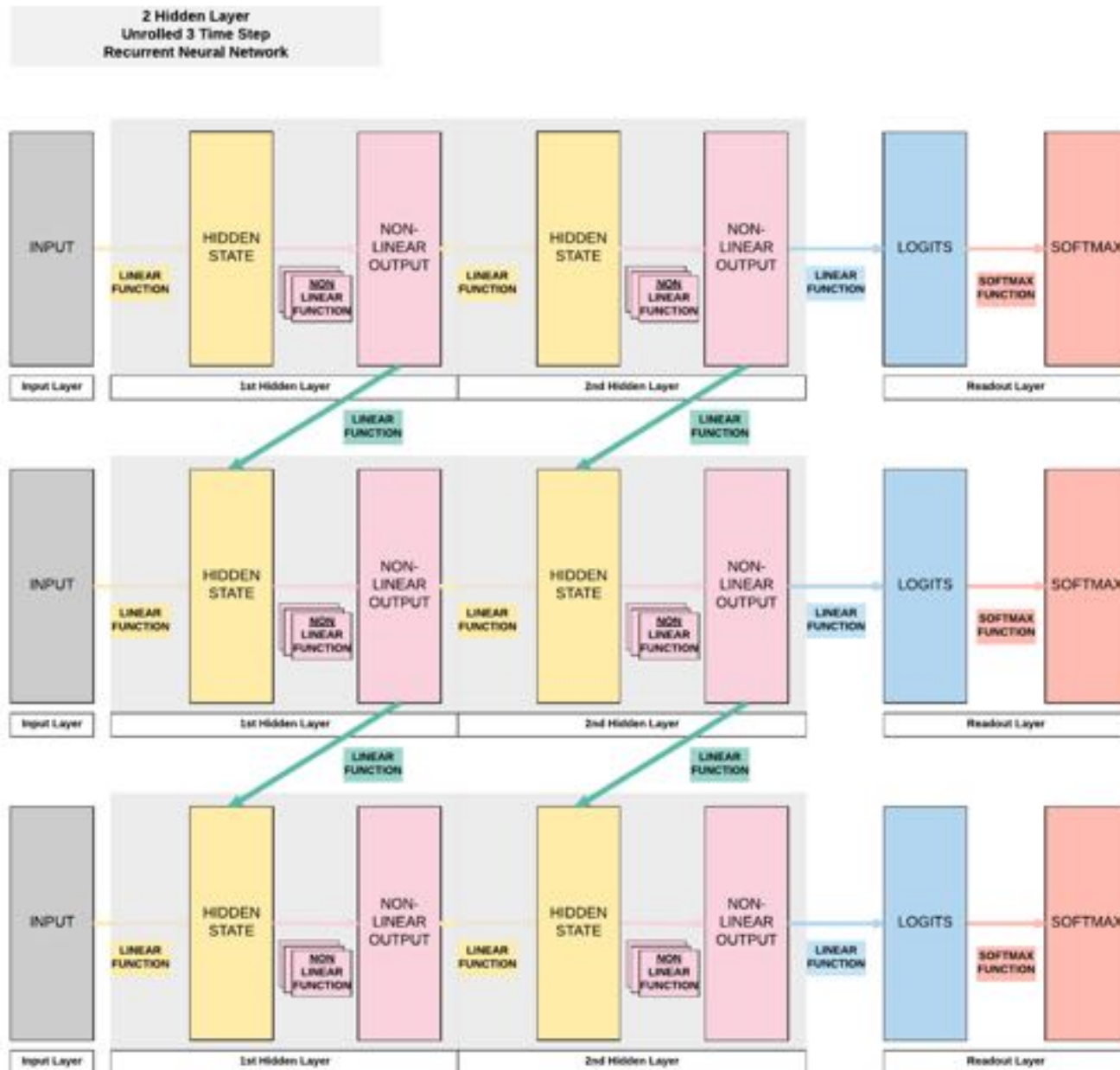
many to many



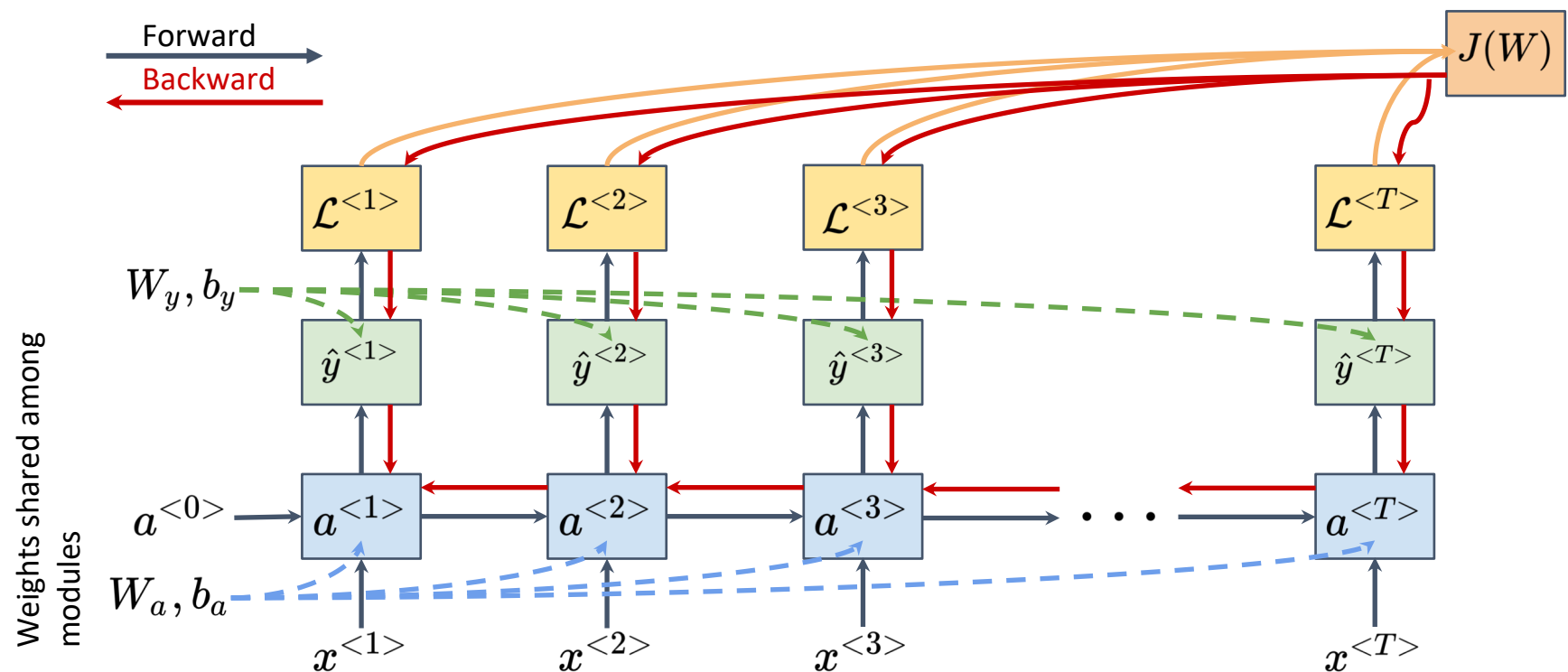
many to many



Deep RNN example



Backpropagation through time!



$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = y^{(i)} \log(f(x^{(i)})) + (1 - y^{(i)}) \log(1 - f(x^{(i)}))$$

Cross-entropy loss for each output

$$J(W) = \sum_{t=1}^T \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

Total cost

RNN problems - vanishing gradients

Long term feature dependencies are very difficult to learn with a standard RNN.

- *"The cat, which already ate ..., **was** full."*
- *"The cats, which already ate ..., **were** full."*

Vanishing/Exploding gradients

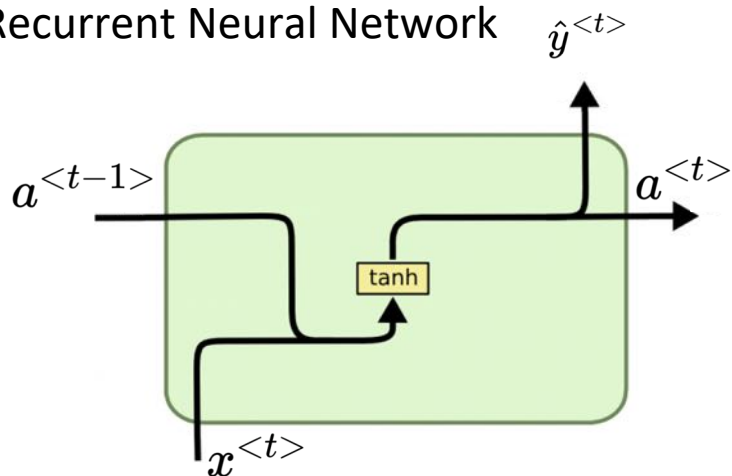
Solutions:

1. Using ReLU activation functions to prevent shrinking the gradients.
2. Initialize the weights to the identity matrix
 - a. Biases still initialized to zero
3. Use a more complex recurrent unit with gates to control what information is passed through.

Gated cells
LSTM, GRU, etc.

RNN

Recurrent Neural Network

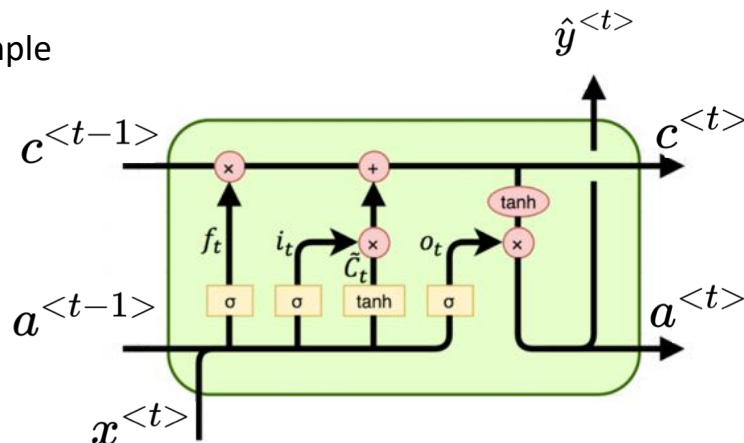
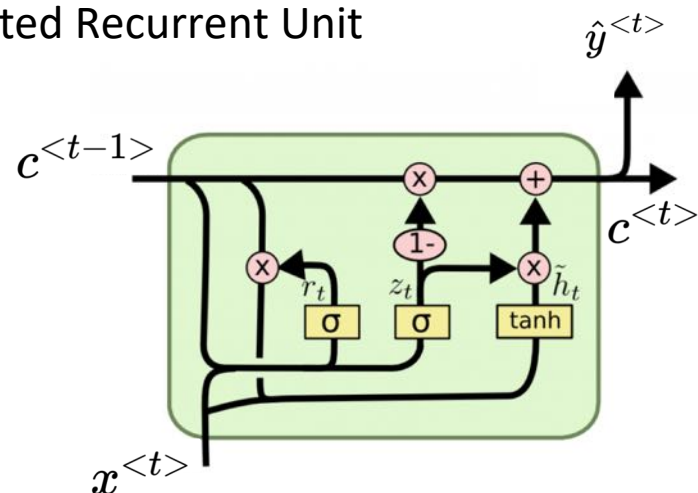


$$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$$

$g() = \tanh$ is the above example

GRU

Gated Recurrent Unit



LSTM

Long-short term memory

Gated Recurrent Unit (GRU)

c = memory cell (remember singular or plural).

$$c^{<t>} = a^{<t>}$$

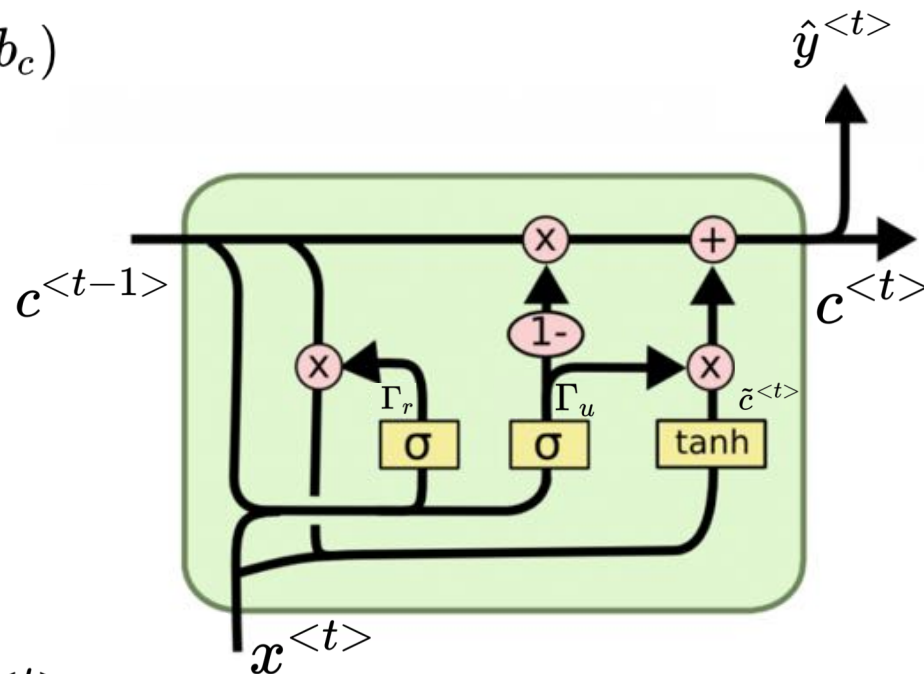
$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^t] + b_u)$$

$$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^t] + b_r)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$$

$\Gamma_u = 1$ $\Gamma_u = 0$ $\Gamma_u = 1$
 "The **cat**, which already ate ..., **was** full."



Update gate: Γ_u decides when to update $c^{<t>}$

Relevance gate: Γ_r how relevant is the last feature in computing the next output?

Intuition: Gates are always on or off. They are modelled with a sigmoid function, so in practice very close to 0 or 1.

* is element-wise vector multiplication.

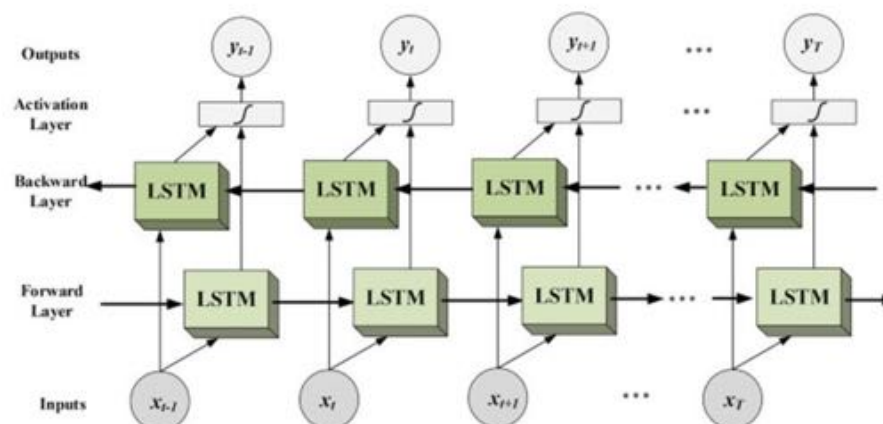
Gated cells problems

Forward propagating a sequence model only uses earlier information to predict context information.

Example, is Teddy a name?

- *He said, "Teddy Roosevelt was a great President."*
- *He said, "Teddy bears are on sale!"*

One solution is to use a *Bidirectional RNN* structure



Still problematic - Words in context

Danish to english translation

- *Hans* is a name but does also mean *his*.
- *Ged* means *goat*, but often used when something went *wrong*.
- *Regner* can be both *rains* and *calculates/computes*.
- *På spanden* literally means *on the bucket* but often used to say *in trouble*.

Goat has gone into the budget
because Hans is raining badly, so in
short he is out on the bucket month.

2019 update



A note on Natural Language Processing

Generalizing from one example of "apple juice" to another of "orange juice" is not more intuitive than to "orange man".

I want a glass of orange _____.

I want a glass of apple _____.

1-hot representation.

Man (5391) Woman (9853) King (4914) Queen (7157) Apple (456) Orange (6257)

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

	Man	Woman	King	Queen	Apple	Orange
Gender	-1.0	1.0	-0.95	0.97	0.0	0.01
Royal	0.0	0.0				
Age						
Food					0.95	0.97
...						

Sequence modelling summary

- RNNs can be used to model sequence tasks.
- Model sequences are modelled via a recurrence relation.
- Training RNNs can be done with back-propagation through time and a gradient based optimizer.
- Gated cells like GRU let us model long-term dependencies.

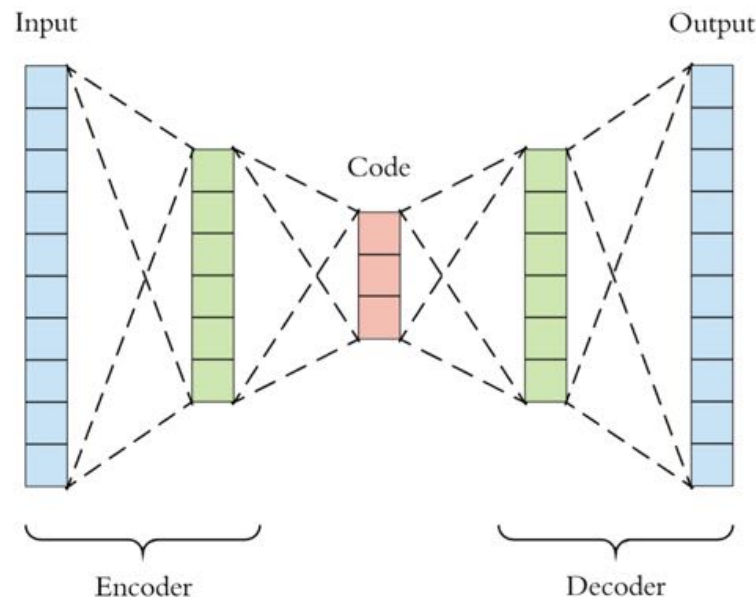
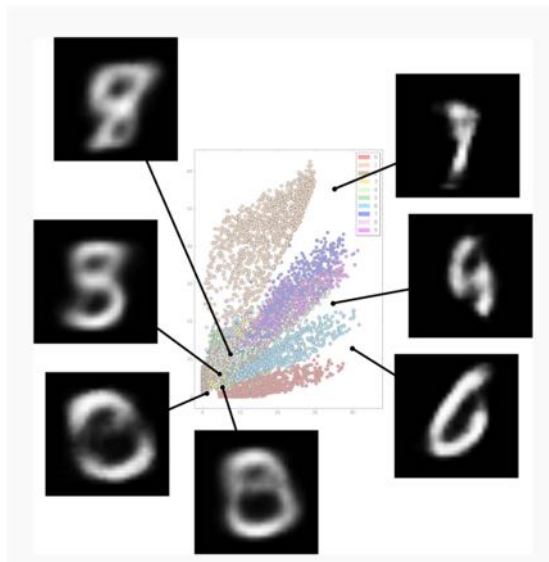
Generative Models

Variational Autoencoders (VAEs)

Generative Adversarial Networks (GANs)

Autoencoders

- Basic Autoencoder network
 - With linear activation functions, this is similar to Principal Component Analysis (PCA) - later topic in this course.



$$\phi : \mathcal{X} \rightarrow \mathcal{F}$$

$$\psi : \mathcal{F} \rightarrow \mathcal{X}$$

$$\phi, \psi = \arg \min_{\phi, \psi} \|X - (\psi \circ \phi)X\|^2$$

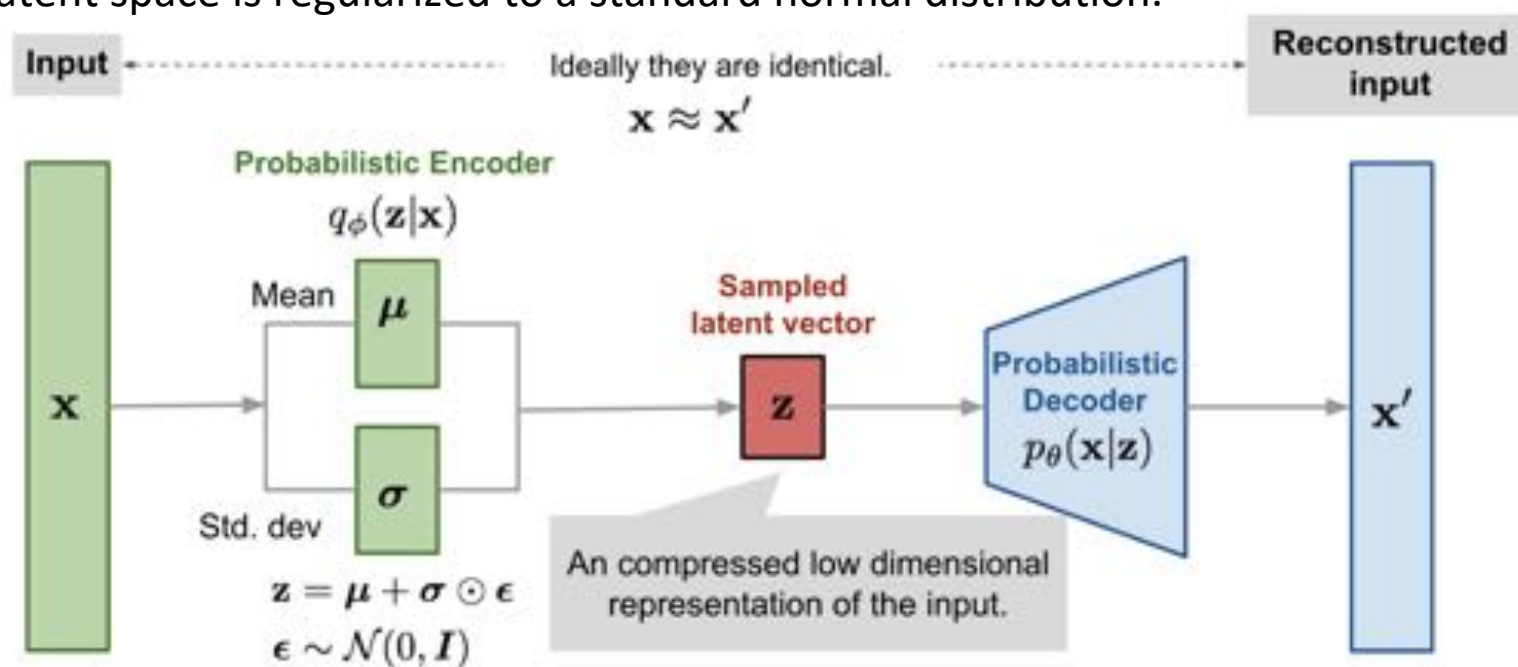
$$\mathcal{L}(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|^2$$

The basic autoencoder contains gaps in the latent space.

Latent space not well separated.

Variational Autoencoders (VAE)

- Encoder is learning an approximation of the posterior distribution.
- Latent space is regularized to a standard normal distribution.

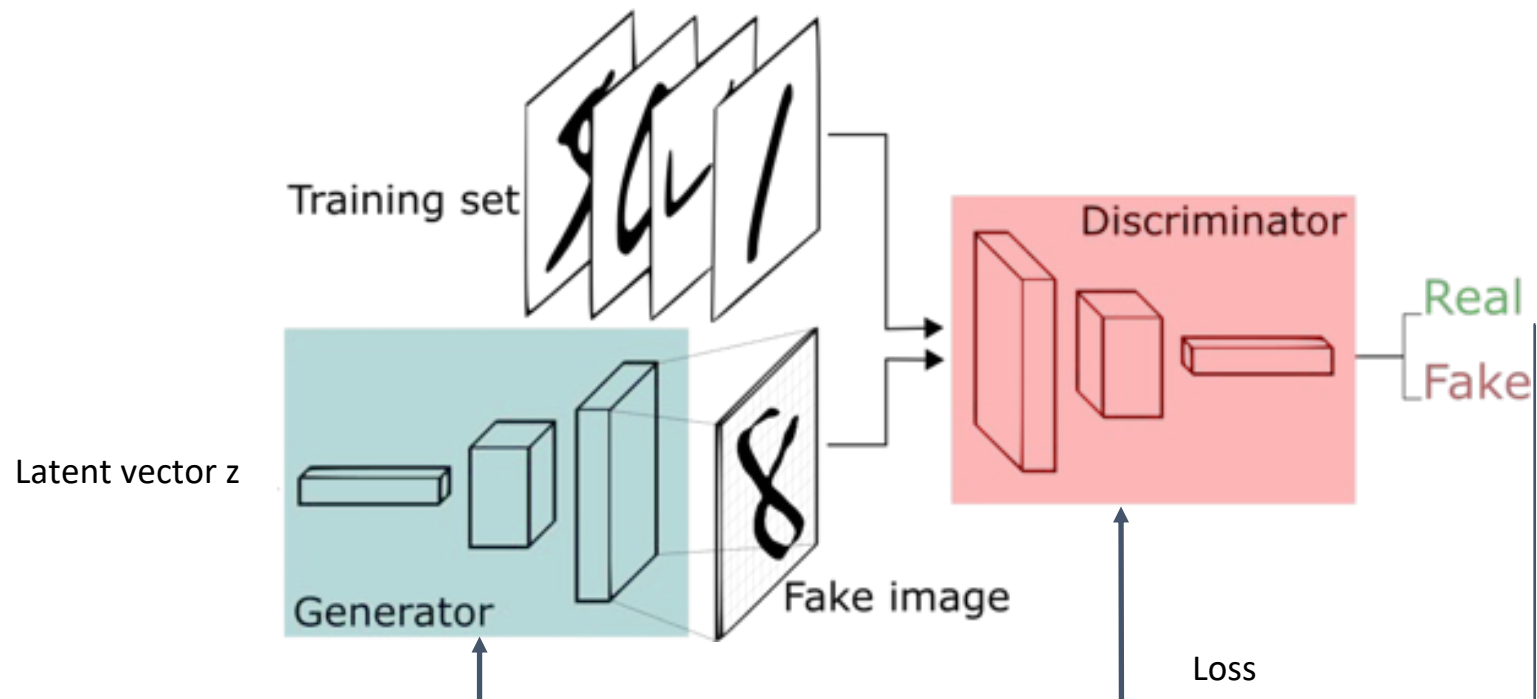


Means of MNIST digits



Generative Adversarial Network (GAN)

- Generator objective: Fool the discriminator network by generating more real images.
- Discriminator objective: Become better in discriminating real and fake images



StyleGAN



Figure 2. Uncurated set of images produced by our style-based generator (config F) with the FFHQ dataset. Here we used a variation of the truncation trick [42, 5, 34] with $\psi = 0.7$ for resolutions $4^2 - 32^2$. Please see the accompanying video for more results.

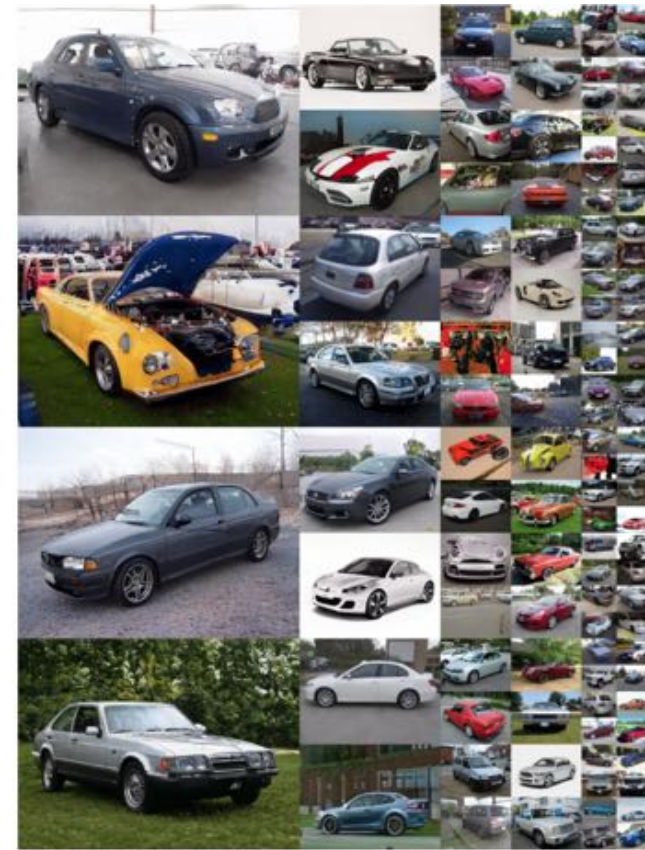


Figure 11. Uncurated set of images produced by our style-based generator (config F) with the LSUN CAR dataset at 512×384 . FID computed for 50K images was 3.27.

<https://thispersondoesnotexist.com/>

<https://thiscatdoesnotexist.com/>

Deep learning outlook

- Deep learning is everywhere and will spread to even more areas in the years to come.
- Even though DL is popular, one should always analyze the problem at hand and pick the best tool.
- Still many open problems within DL:
 - Architecture understanding.
 - Reasoning capabilities, e.g. from context.
 - Robustness against adversarial attacks.
 - Fully unsupervised learning systems to avoid tedious labeling process.

Deep learning energy

Training artificial intelligence is an **energy intensive process**. New estimates suggest that the carbon footprint of training a single AI is as much as 284 tonnes of carbon dioxide equivalent – **five times the lifetime emissions of an average car**.

THE DAILY NEWSLETTER
Sign up to our daily email newsletter

NewScientist

News **Technology** Space Physics Health Environment Mind Video | Tours Events Jobs

Creating an AI can be five times worse for the planet than a car



TECHNOLOGY 6 June 2019

By Donna Lu



Server farms are used to train AIs
Tommy Lee Walker/Getty

Deep learning detection robustness

Spot a pedestrian walking walking in front of a car coming with 20 MPH.

Only 40% of adult collisions in optimal conditions were avoided.

At night, the systems didn't even ping the driver to reduce speed.

AAA Car Testing Shows Pedestrian Detection Tech Is Far From Effective

By Ryan Whitwam on October 7, 2019 at 11:02 am [Comment](#)



Credits

Books:

- <https://www.deeplearningbook.org/>
- <http://neuralnetworksanddeeplearning.com/>

Online Course from MIT:

- <http://introtodeeplearning.com/>

Online course from Stanford University:

- <https://www.coursera.org/specializations/deep-learning?>

Other

- cs231n.github.io
- appliedgo.net
- brohrer.github.io
- learnopencv.com