Neural Machine Translation Techniques Used By Google

Sophia Mitchellette

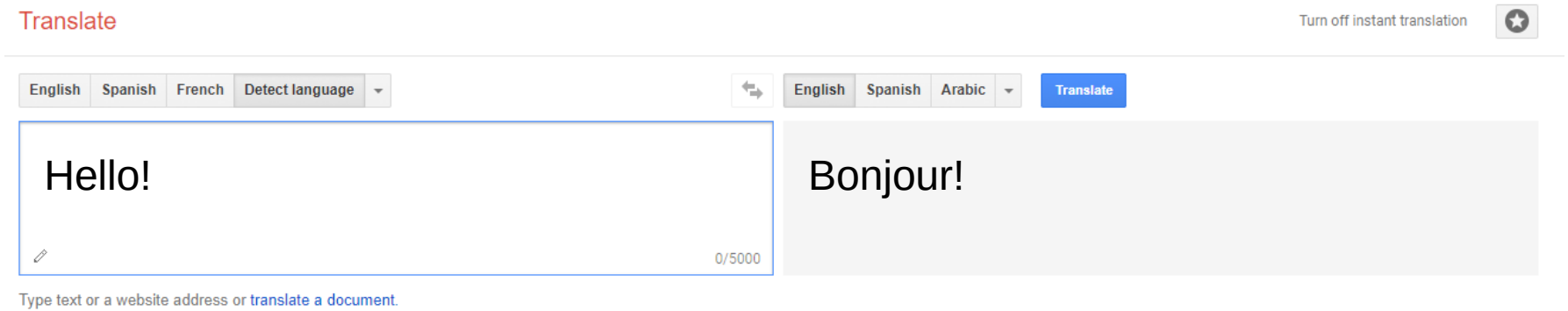
Outline

- Introduction
- Neural Networks
- Residual Connections
- Attention Network
- Summary

Outline

- Introduction
  - What Is Machine Translation?
  - Human Translation vs. Machine Translation
  - How Machine Translation Systems See Words
- Neural Networks
- Residual Connections
- Attention Network
- Summary

Introduction - What Is Machine Translation?



Machine Translation - translation from one language to another performed by a machine instead of a human

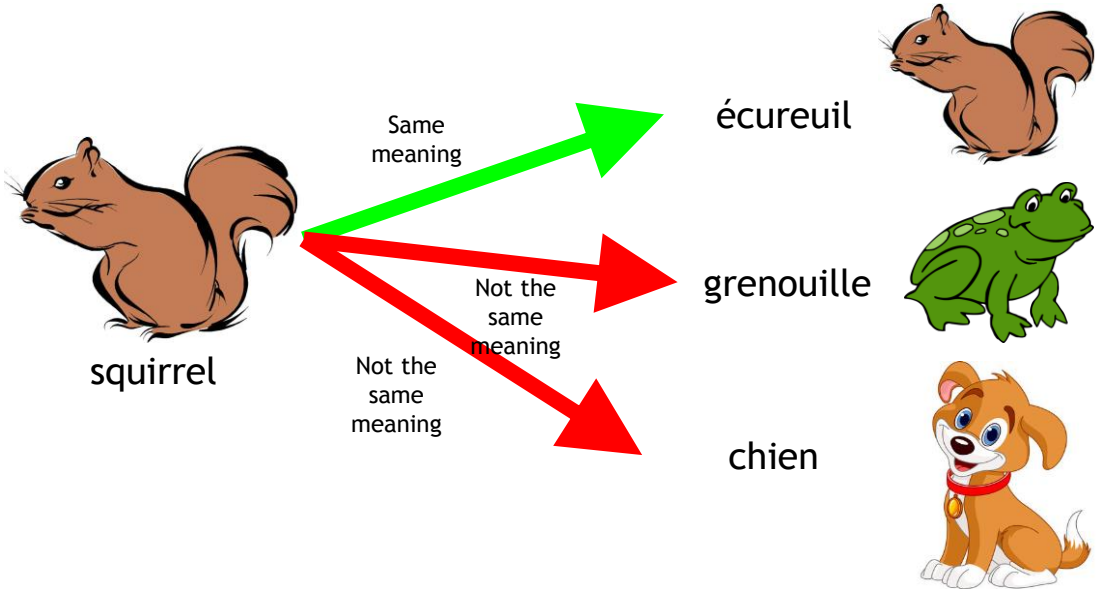
Google Translate

Google's Machine Translation Service

Behind the scenes: Google's Neural Machine Translation System (GMNT)

Over 500 million people use every day

Introduction - Human Translation

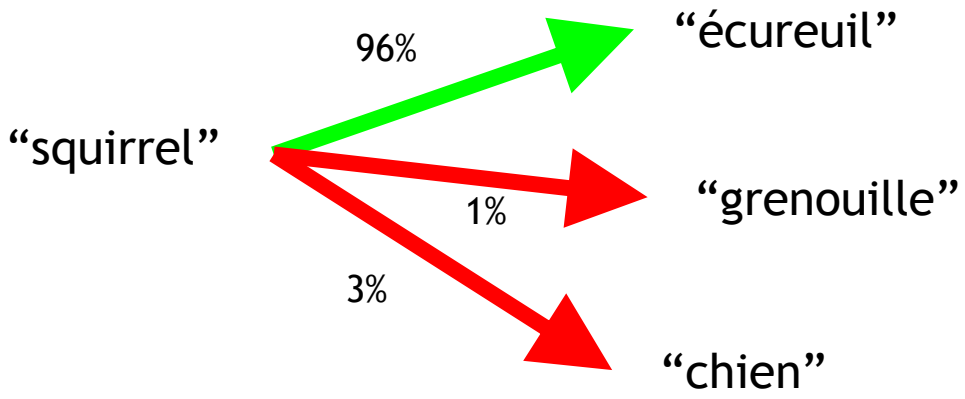


- Humans translate by finding words, sentences and phrases that have the same meaning in both languages.

http://www.clipartpanda.com/clipart_images/squirrel-clip-art-hd-5521910

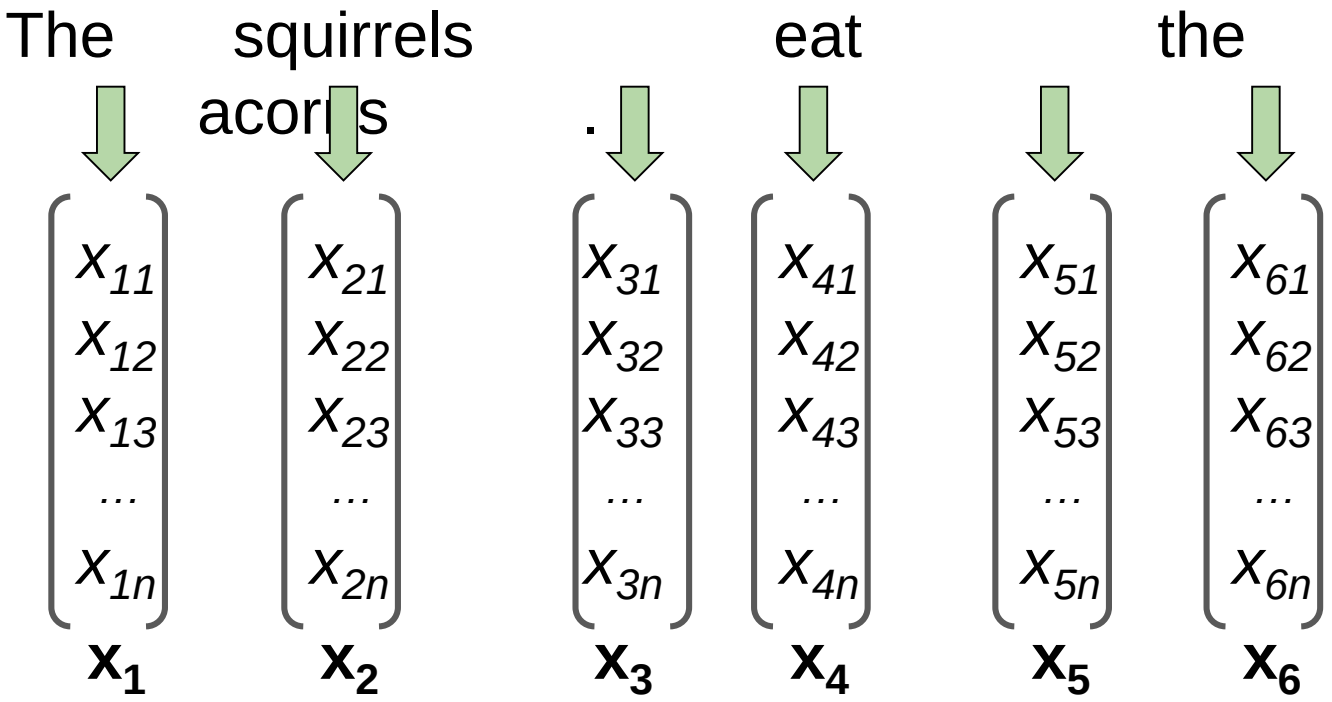
http://clipartix.com/frog-clip-art-image-10263/

Introduction - Machine Translation



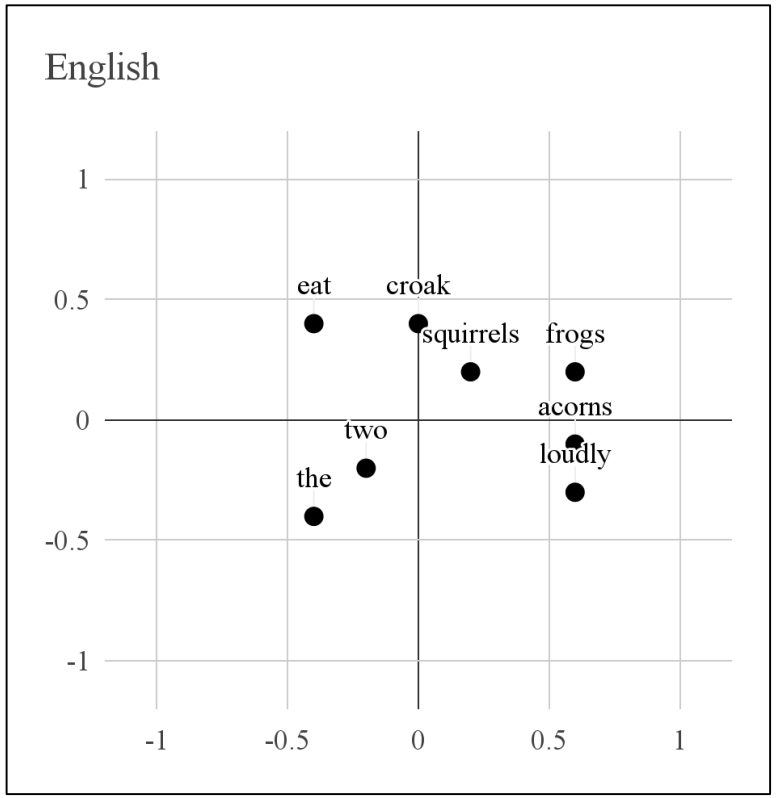
- Look at statistical probabilities for the best translation
- Words and sentences are not a communication of meaning

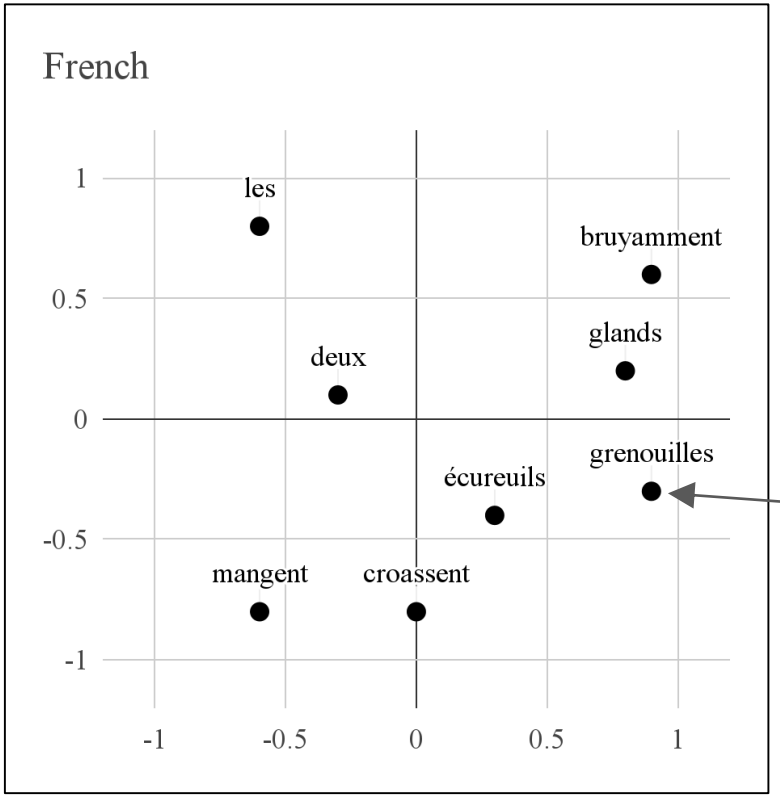
Introduction - How Machine Translation Systems See Words



- Words and sentences are vectors.
- *Word segmentation* - each word is represented as one vector

Introduction - How Machine Translation Systems See Words





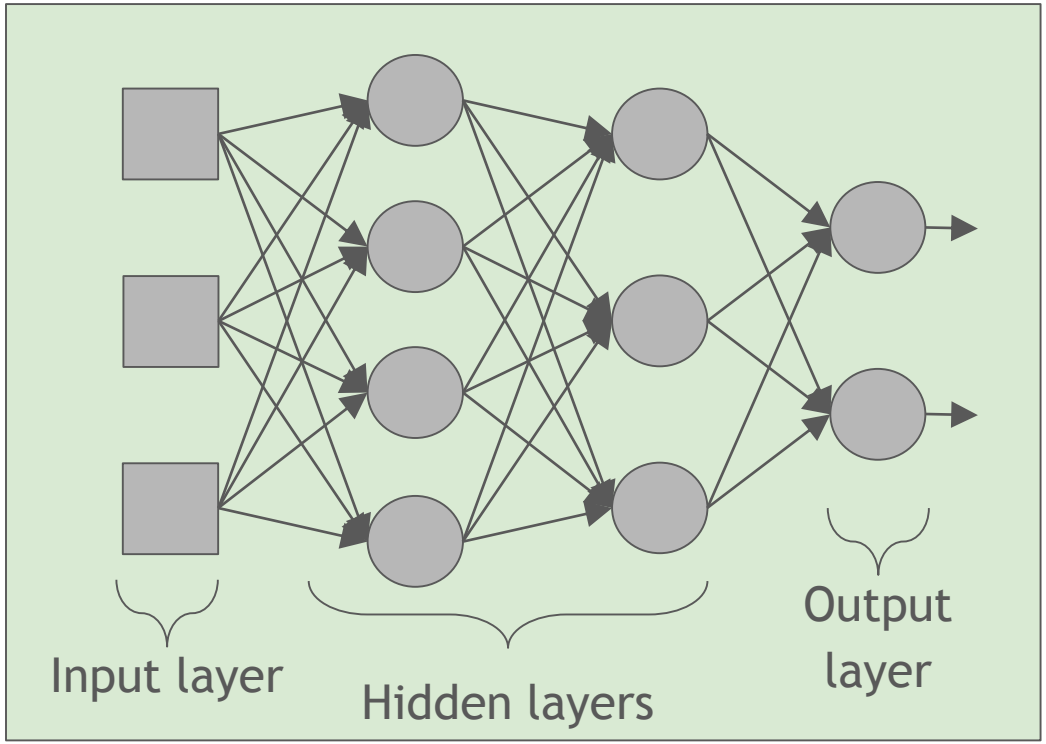
- The points on the coordinate plane represent two-dimensional word vectors.

Outline

- Introduction
- Neural Networks
  - An Overview
  - The Structure of a Node
  - A Neural Network
  - Activation Functions
  - Training a Neural Network
  - Training Error vs. Testing Error
- Residual Connections
- Attention Network
- Summary

Neural Networks - An Overview

- *Nodes* - the building blocks of neural networks.
- Nodes *map* inputs to outputs.
- *mapping* = *function*

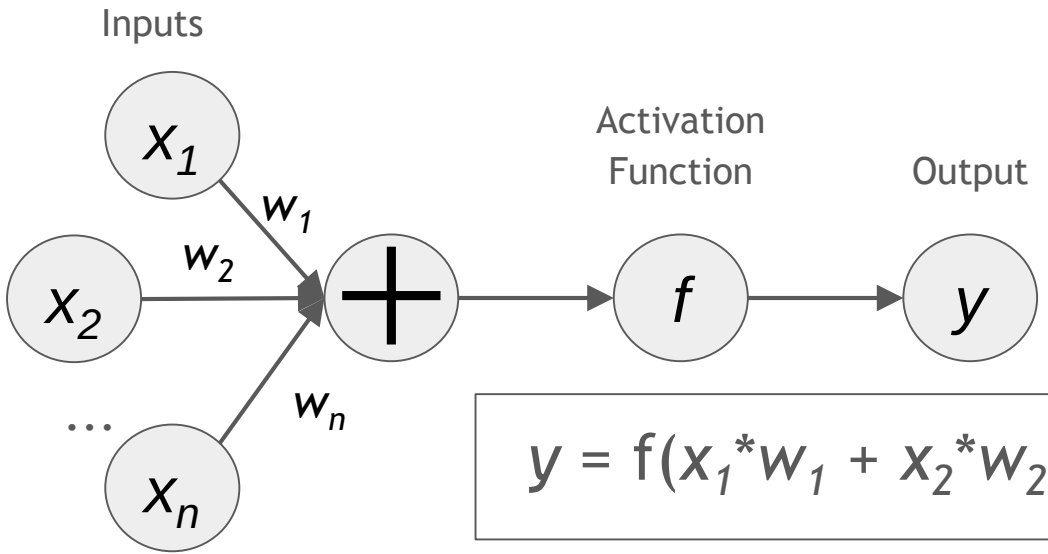


Pictured right: A neural network. Each gray circle is a node. Gray squares are inputs.

Neural Networks - The Structure of a Node

Steps of a node:

1. Inputs are multiplied by their weights.
2. Weighted inputs are summed.
3. Summation is run through activation function.
4. Result of activation function is the output.

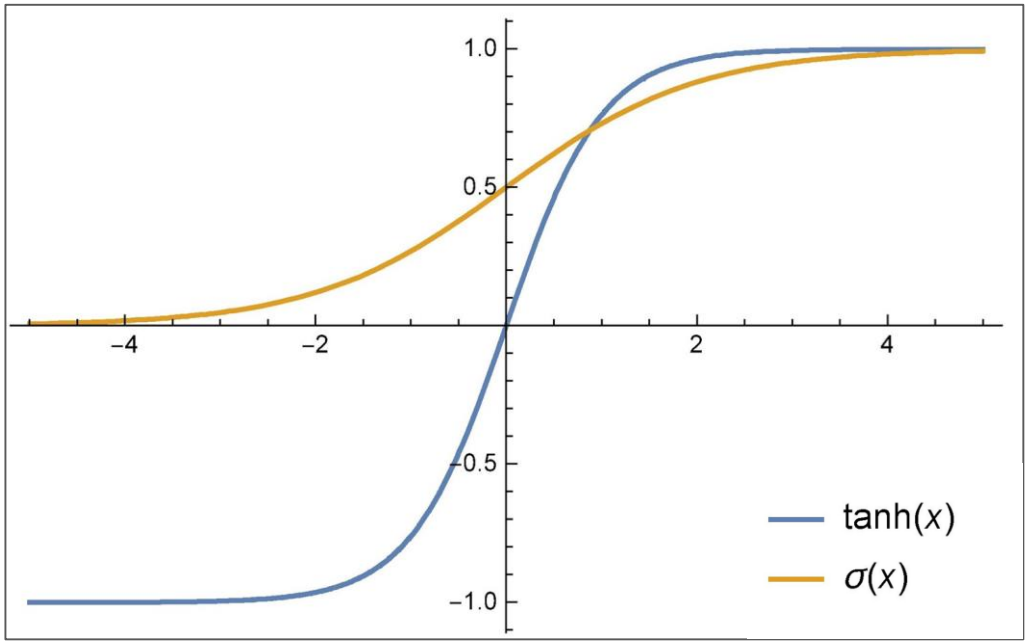


$$y = f(x_1*w_1 + x_2*w_2 + \ldots + x_n*w_n)$$

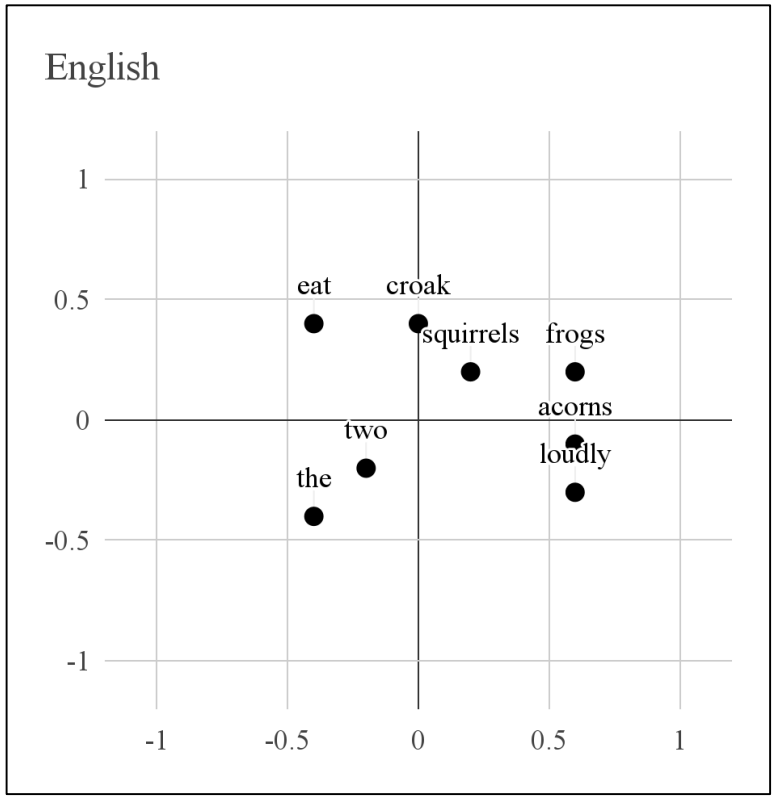
Neural Networks - Activation Functions

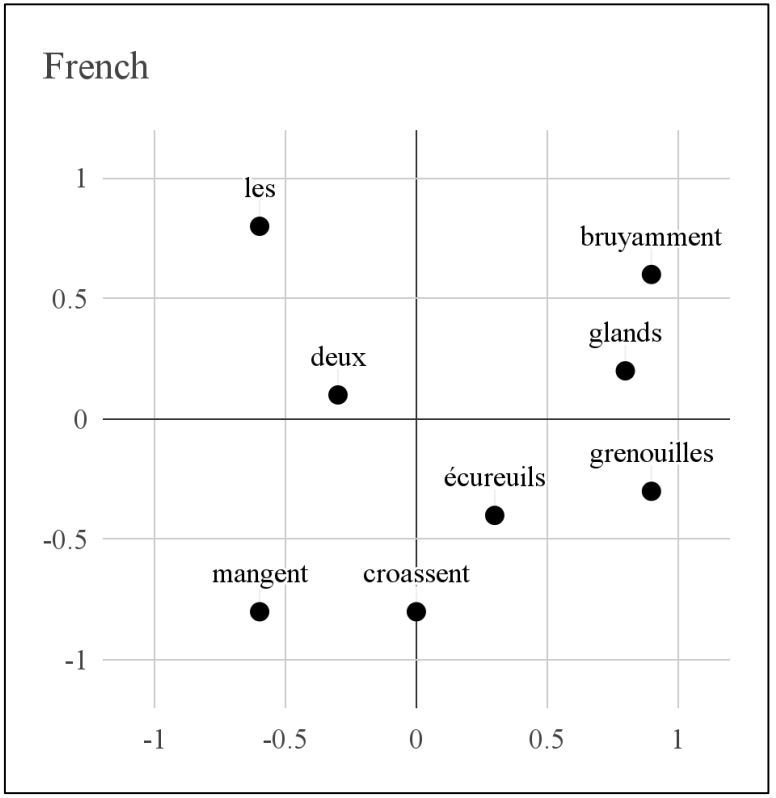
- Need non-linear activation functions for more complex data patterns.

- Hyperbolic tangent (*tanh)* and sigmoid (σ) scale values.



Neural Networks - A Neural Network



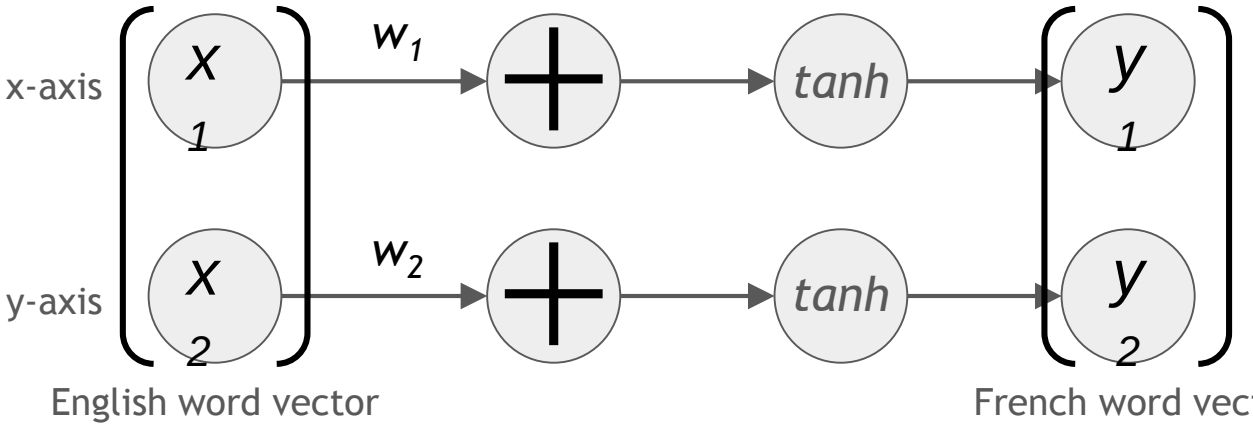


- Our neural network will map from the English words, to their French counterparts.

Neural Networks - A Neural Network

- Network has two nodes
- Takes in English word
- Outputs French word

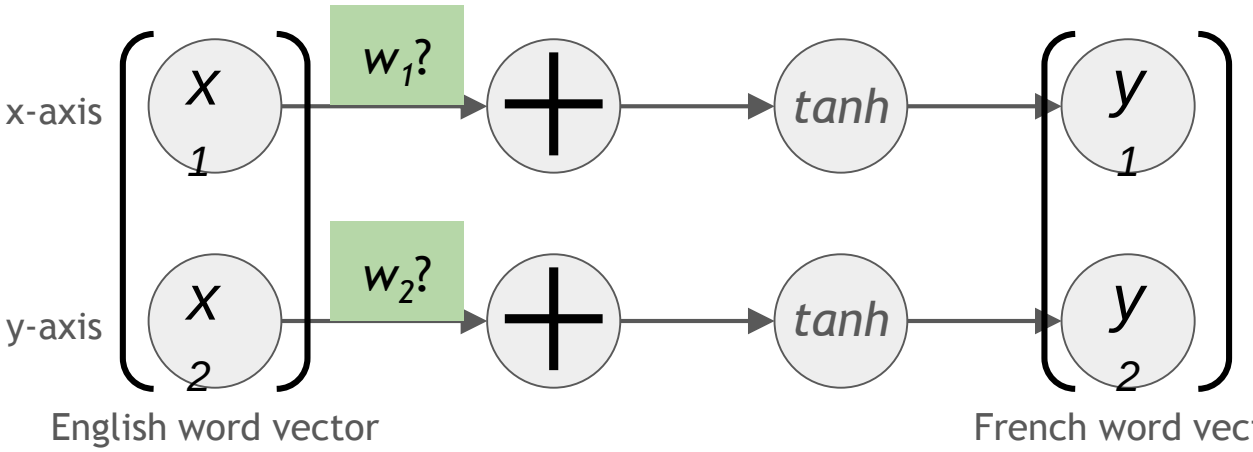
$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} tanh(x_1 * w_1) \\ tanh(x_2 * w_2) \end{pmatrix}$$



Neural Networks - A Neural Network

- Network has two nodes
- Takes in English word
- Outputs French word

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} tanh(x_1 * w_1) \\ tanh(x_2 * w_2) \end{pmatrix}$$



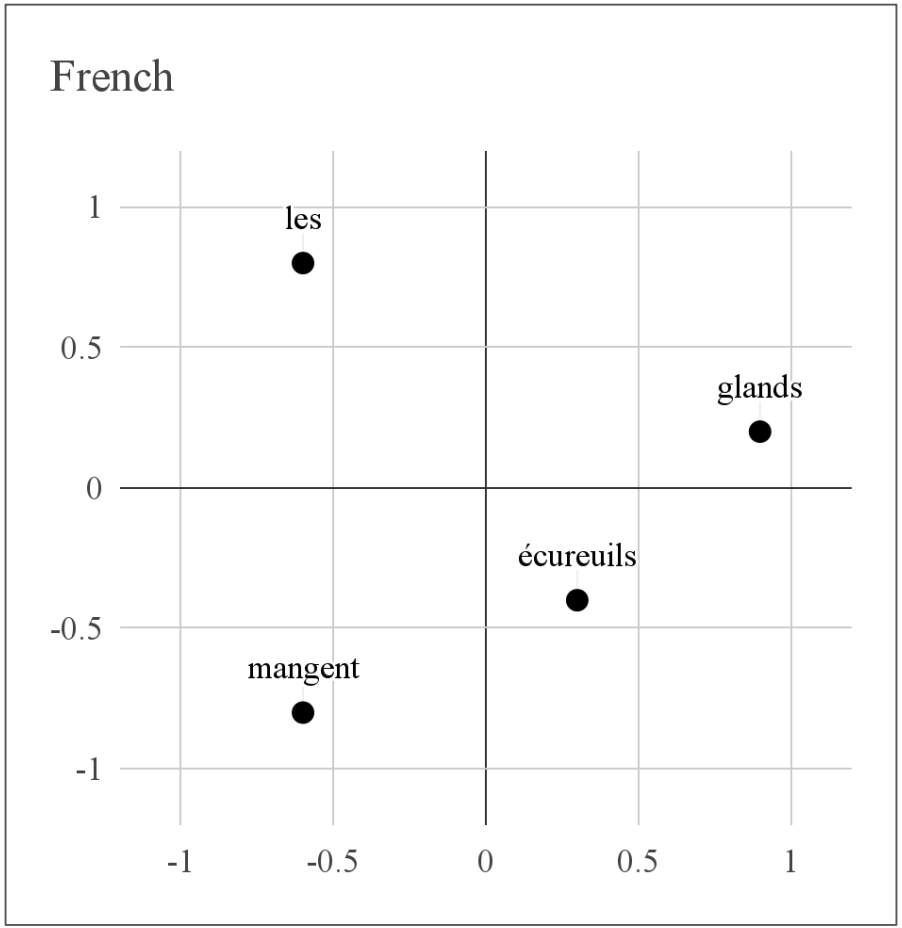
Neural Networks - Training a Neural Network

- Neural network is given examples of what the output should be for given inputs.

English Word	English Word Vector	French Word Vector	French Word
squirrel	$\begin{bmatrix} 0.2 \\ 0.2 \end{bmatrix}$	$\begin{bmatrix} 0.3 \\ -0.4 \end{bmatrix}$	écureuils
acorns	$\begin{bmatrix} 0.6 \\ -0.4 \end{bmatrix}$	$\begin{bmatrix} 0.9 \\ 0.8 \end{bmatrix}$	glands
eat	$\begin{bmatrix} -0.2 \\ 0.4 \end{bmatrix}$	$\begin{bmatrix} -0.3 \\ -0.8 \end{bmatrix}$	mangent
the	$\begin{bmatrix} -0.4 \\ -0.4 \end{bmatrix}$	$\begin{bmatrix} -0.6 \\ 0.8 \end{bmatrix}$	les
	Inputs	Desired Outputs	

Neural Networks - Training a Neural Network

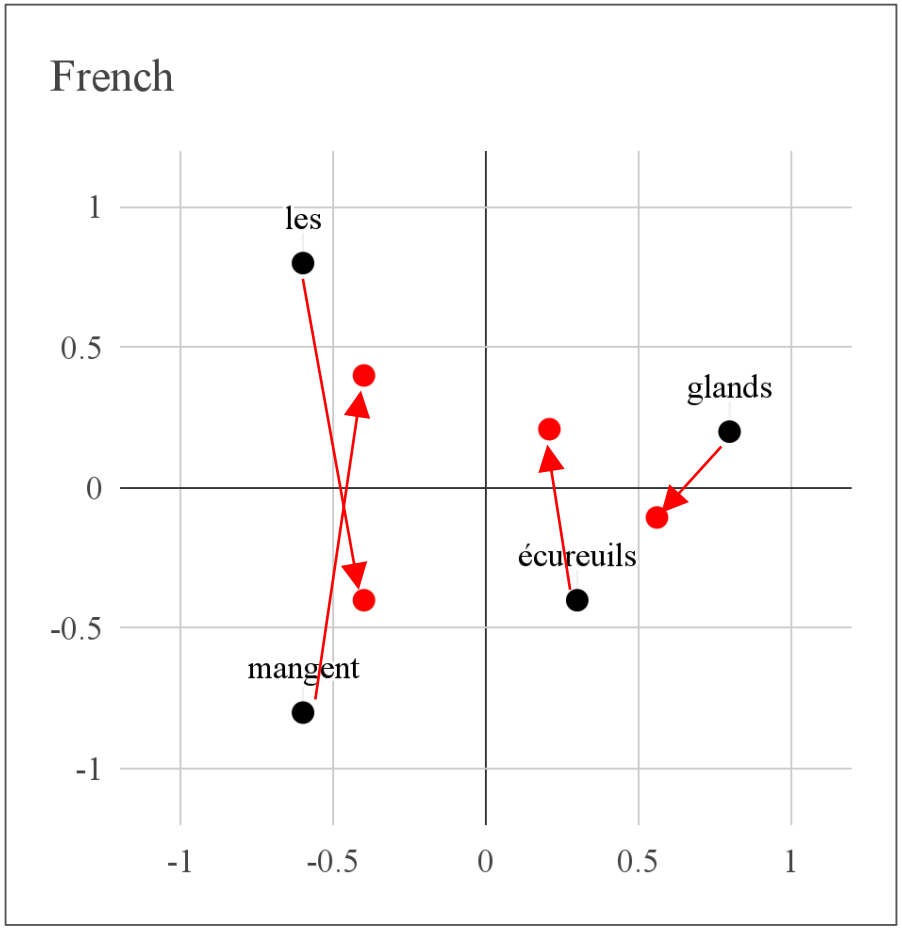
- Goal: inputs map to desired vectors
- Network starts with initial weights
- Error evaluated. New weights tried.
- *Iteration* - an attempt



Iteration	w_1	w_2
1		
2		
3		
4		

Neural Networks - Training a Neural Network

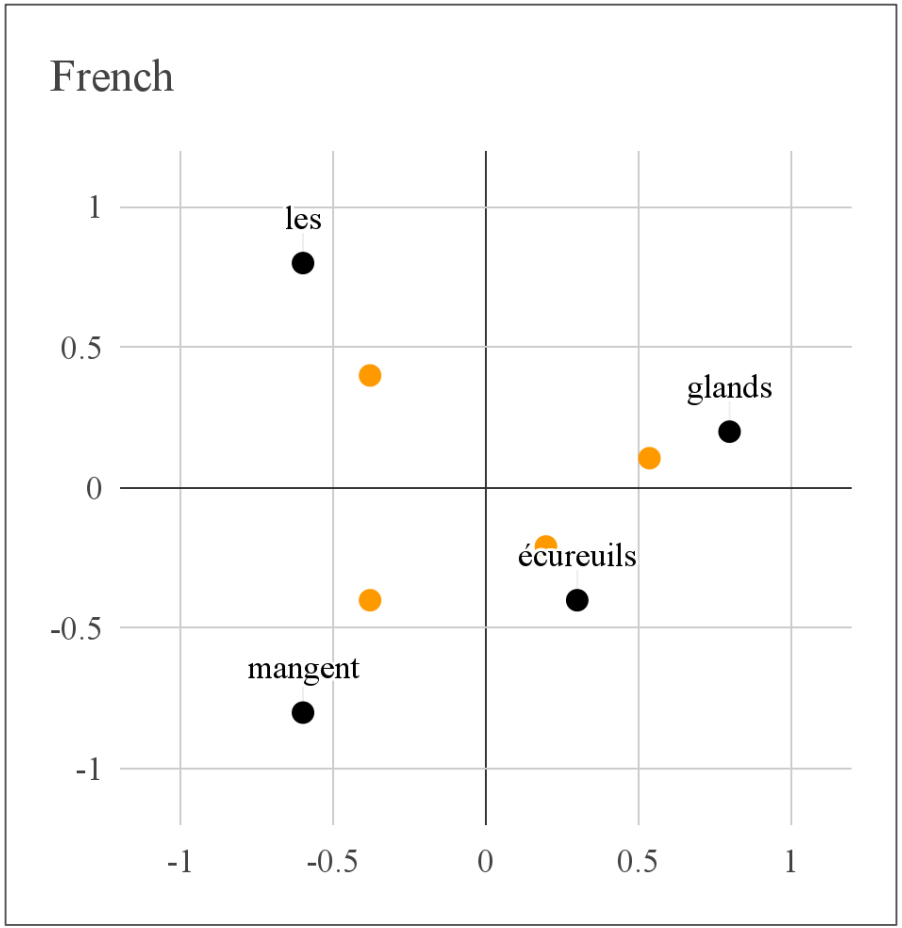
- Goal: inputs map to desired vectors
- Network starts with initial weights
- Error evaluated. New weights tried.
- *Iteration* - an attempt



Iteration	w_1	w_2
1	1.059	1.059
2		
3		
4		

Neural Networks - Training a Neural Network

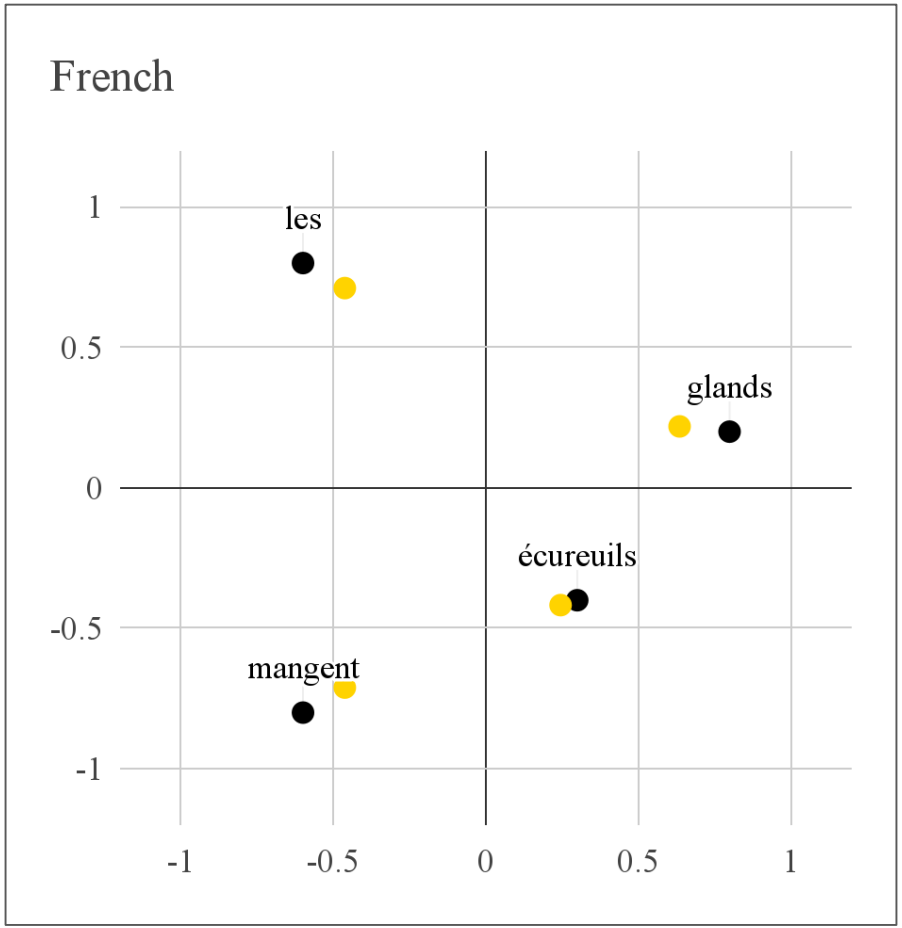
- Goal: inputs map to desired vectors
- Network starts with initial weights
- Error evaluated. New weights tried.
- *Iteration* - an attempt



Iteration	w_1	w_2
1	1.059	1.059
2	1	-1.059
3		
4		

Neural Networks - Training a Neural Network

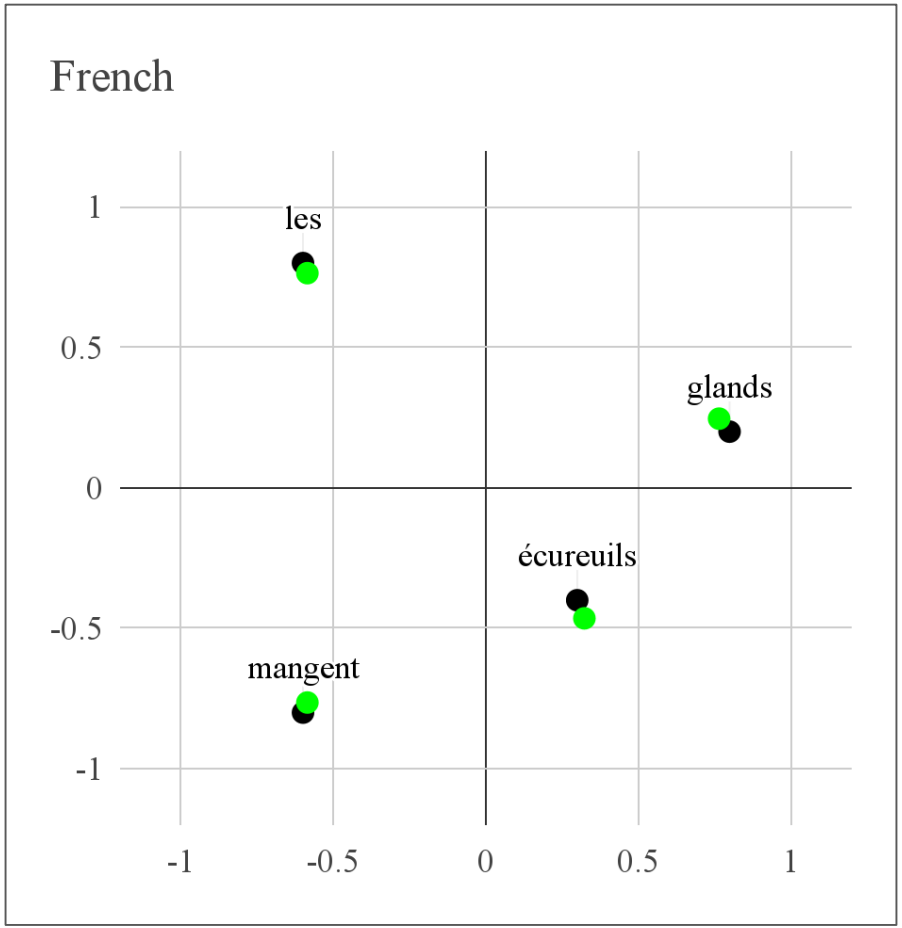
- Goal: inputs map to desired vectors
- Network starts with initial weights
- Error evaluated. New weights tried.
- *Iteration* - an attempt



Iteration	w_1	w_2
1	1.059	1.059
2	1	-1.059
3	1.252	-2.222
4		

Neural Networks - Training a Neural Network

- Goal: inputs map to desired vectors
- Network starts with initial weights
- Error evaluated. New weights tried.
- *Iteration* - an attempt

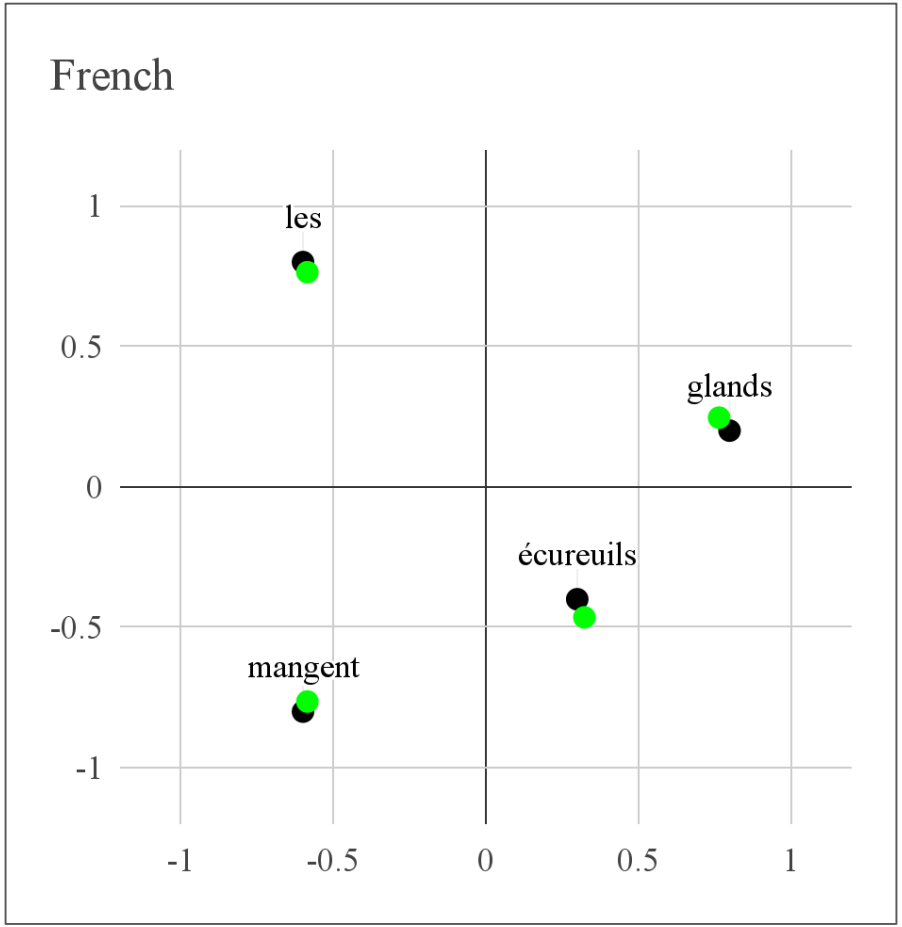


Iteration	w_1	w_2
1	1.059	1.059
2	1	-1.059
3	1.252	-2.222
4	1.683	-2.515

Neural Networks - Testing Error vs. Training Error

- *Training error* - the network's error when run over the data it was trained on.

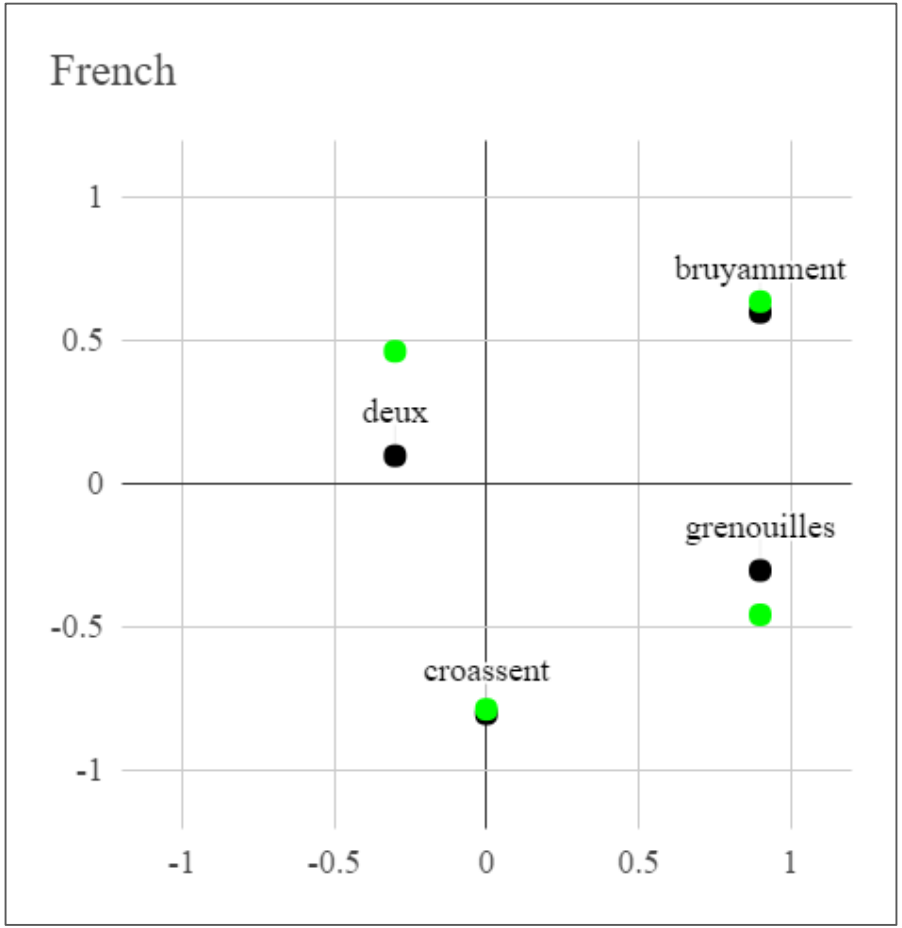
English Word	English Word Vector	French Word Vector	French Word
squirrel	$\begin{pmatrix}0.2\\0.2\end{pmatrix}$	$\begin{pmatrix}0.3\\-0.4\end{pmatrix}$	écureuils
acorns	$\begin{pmatrix}0.6\\-0.4\end{pmatrix}$	$\begin{pmatrix}0.9\\0.8\end{pmatrix}$	glands
eat	$\begin{pmatrix}-0.2\\0.4\end{pmatrix}$	$\begin{pmatrix}-0.3\\-0.8\end{pmatrix}$	mangent
the	$\begin{pmatrix}-0.4\\-0.4\end{pmatrix}$	$\begin{pmatrix}-0.6\\0.8\end{pmatrix}$	les



Neural Networks - Testing Error vs. Training Error

- *Testing error* - the network's error when run over new and unfamiliar data.

English Word	English Word Vector	French Word Vector	French Word
frogs	$\begin{pmatrix}0.6\\0.2\end{pmatrix}$	$\begin{pmatrix}0.9\\-0.3\end{pmatrix}$	grenouilles
loudly	$\begin{pmatrix}0.6\\-0.3\end{pmatrix}$	$\begin{pmatrix}0.9\\0.6\end{pmatrix}$	bruyamment
croak	$\begin{pmatrix}0\\-0.4\end{pmatrix}$	$\begin{pmatrix}0\\-0.8\end{pmatrix}$	croassent
two	$\begin{pmatrix}-0.2\\-0.2\end{pmatrix}$	$\begin{pmatrix}-0.3\\0.1\end{pmatrix}$	deux



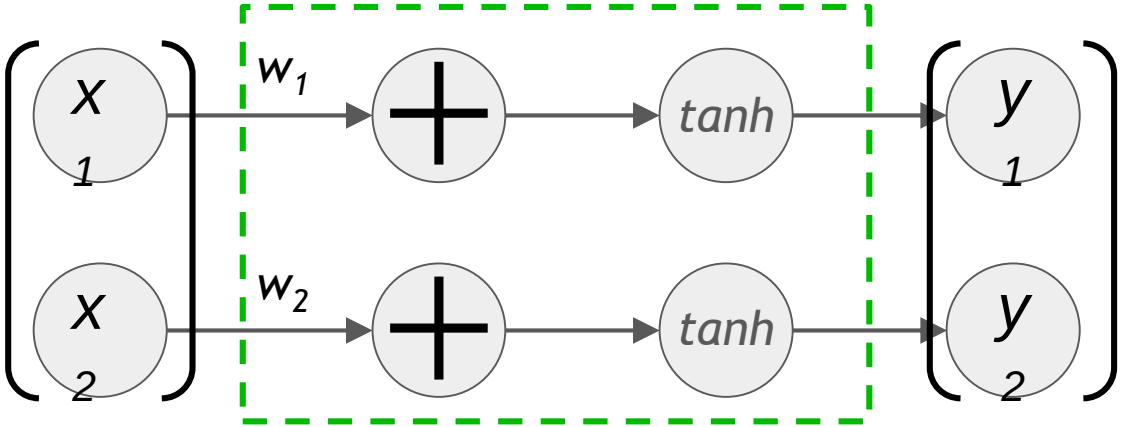
Outline

- Introduction
- Neural Networks
- Residual Connections
  - Definitions
  - Multi-layer networks
  - Difficulty of Training Deep Neural Networks
  - Plain vs. Residual Mapping
  - Evaluations
- Attention Network
- Summary

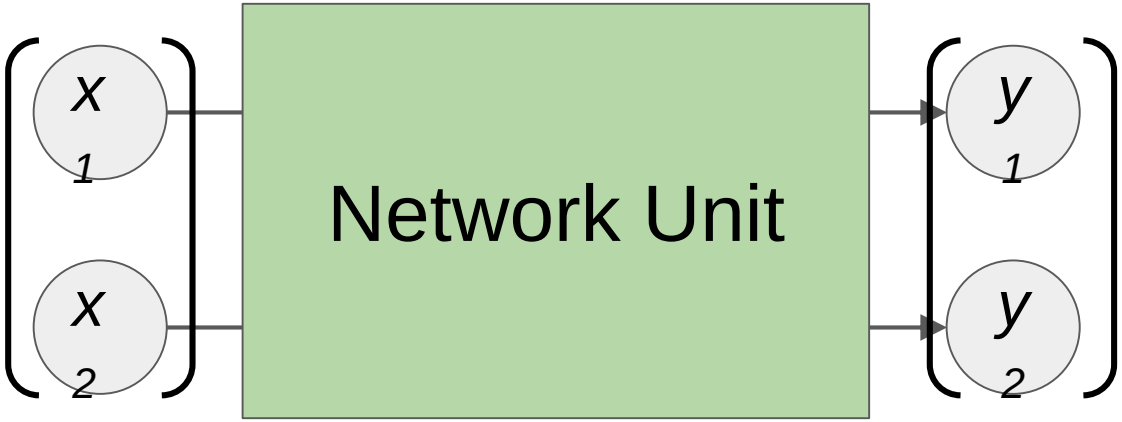
Residual Connections - Plain vs. Residually Connected Networks

- *Plain Neural Network* - does not have residual connections
- *Residually Connected Neural Network* - has residual connections

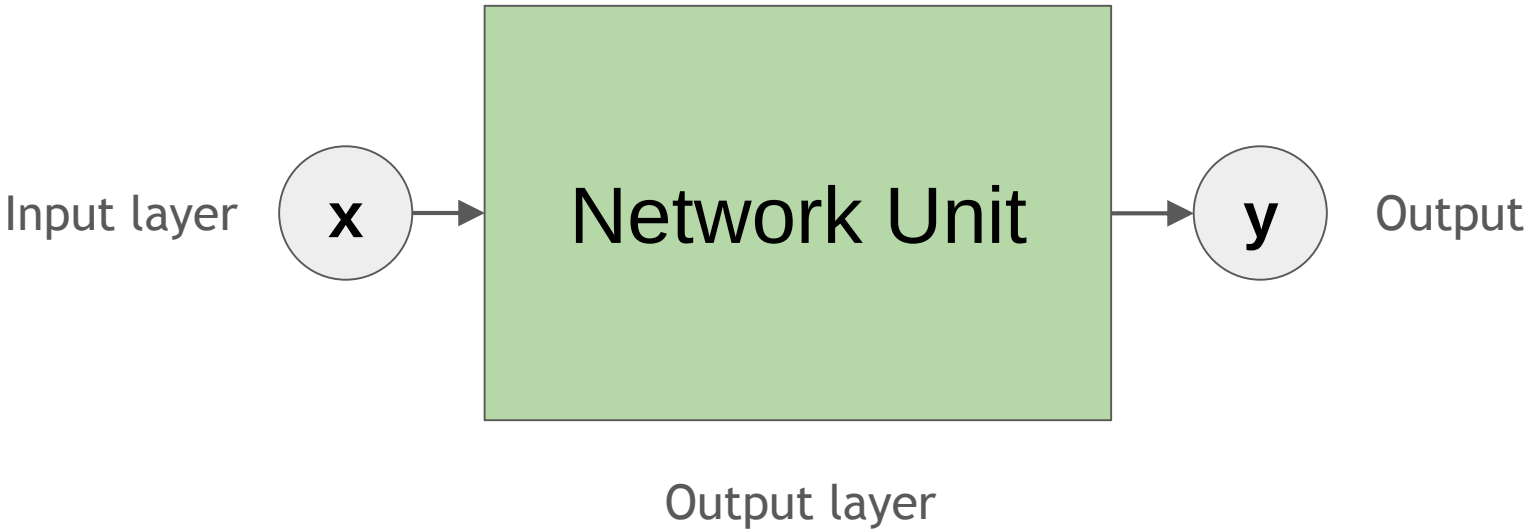
Residual Connections - Multi-layer networks



Residual Connections - Multi-layer networks

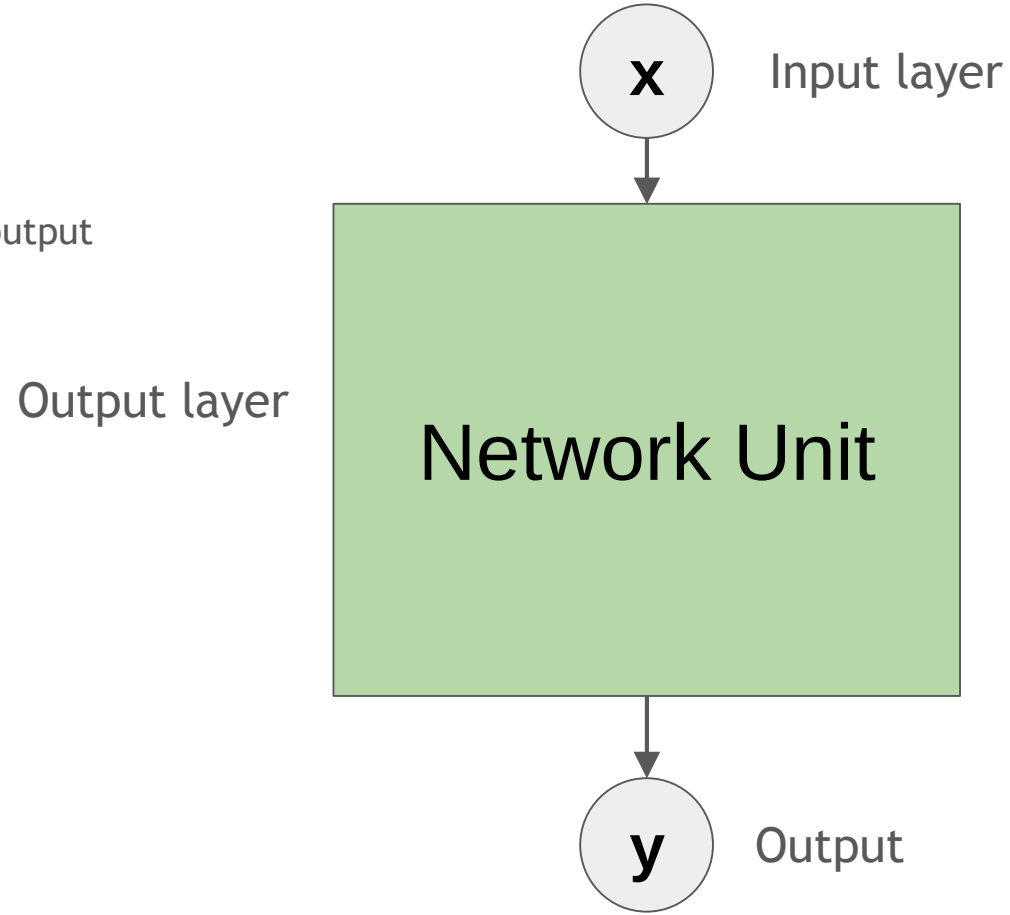


Residual Connections - Multi-layer networks



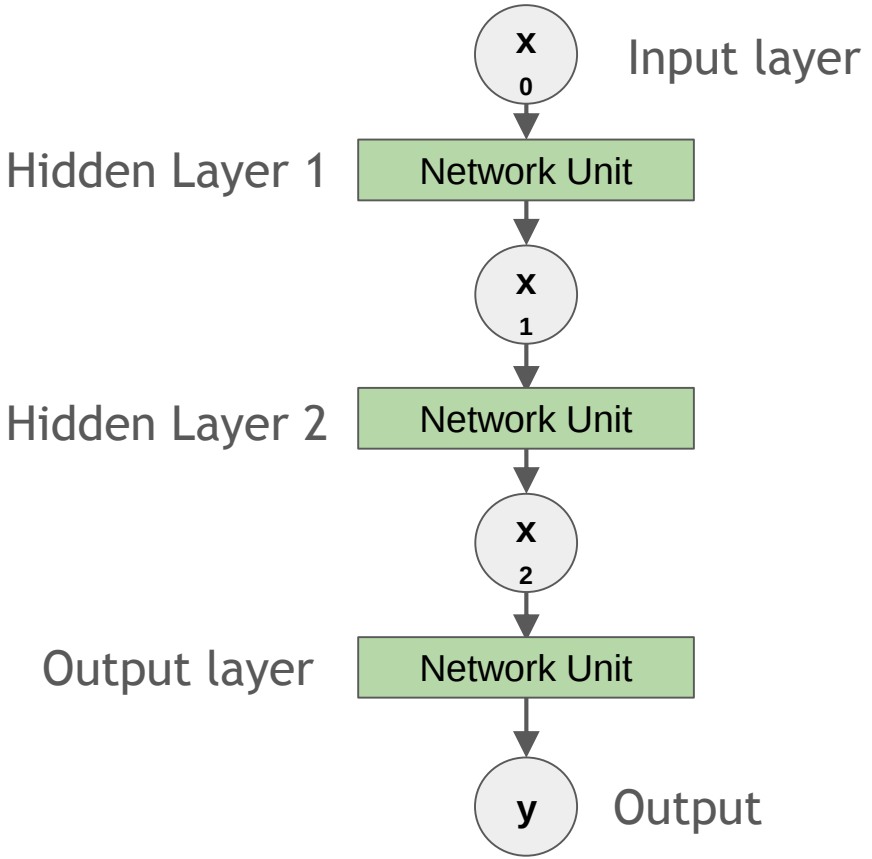
Residual Connections - Multi-layer networks

- In a single-layer network:
  - output layer input = input layer
  - output layer output = network output



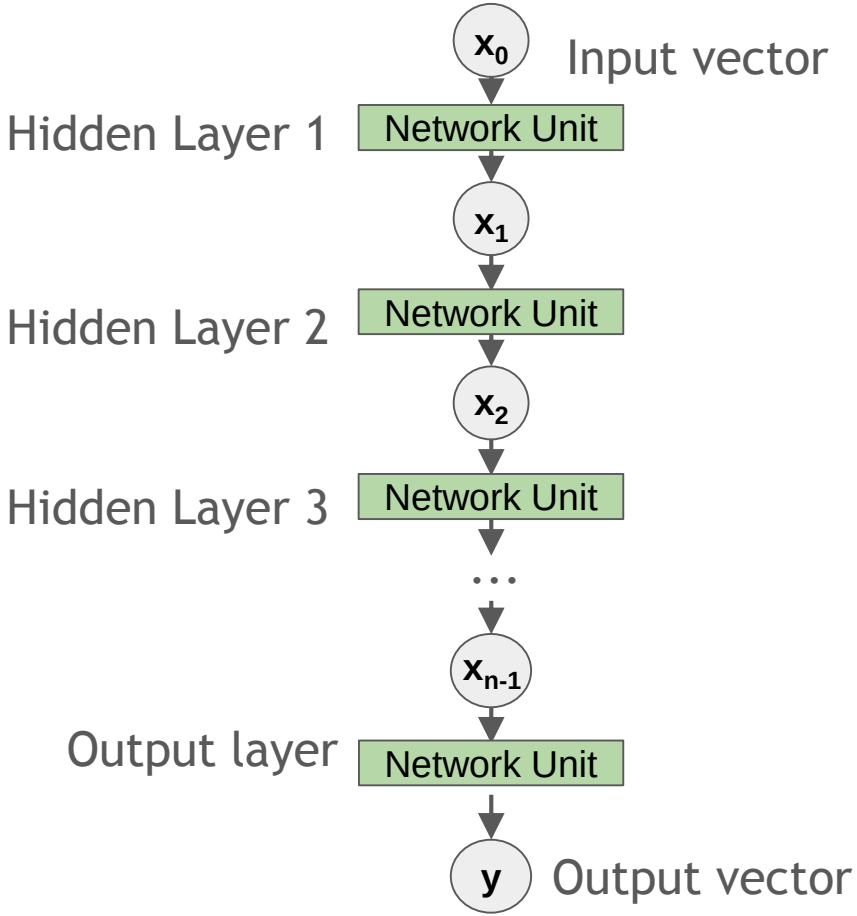
Residual Connections - Multi-layer networks

- In a plain multi-layer network:
  - 1st hidden layer input = input layer
  - 1st h. layer output = 2nd h. layer input
  - etc...
  - output layer output = network output

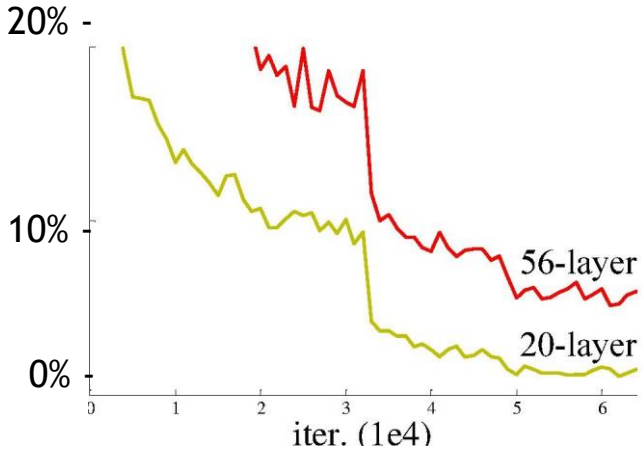


Residual Connections - The Difficulty of Training Plain DNN

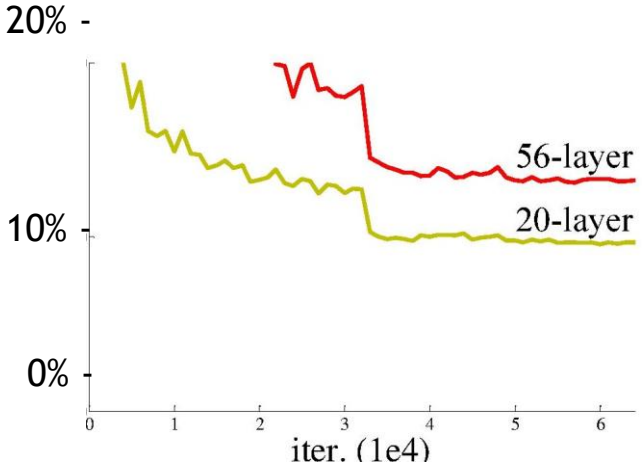
- Deep Neural Networks (DNN)
  - More than one hidden layer
  - More accurate
  - Can account for more complex data patterns.
- Plain DNNs are more difficult to train.
- Residually connected DNNs have residual connections to help with training.



Residual Connections - The Difficulty of Training Plain DNN



Training error

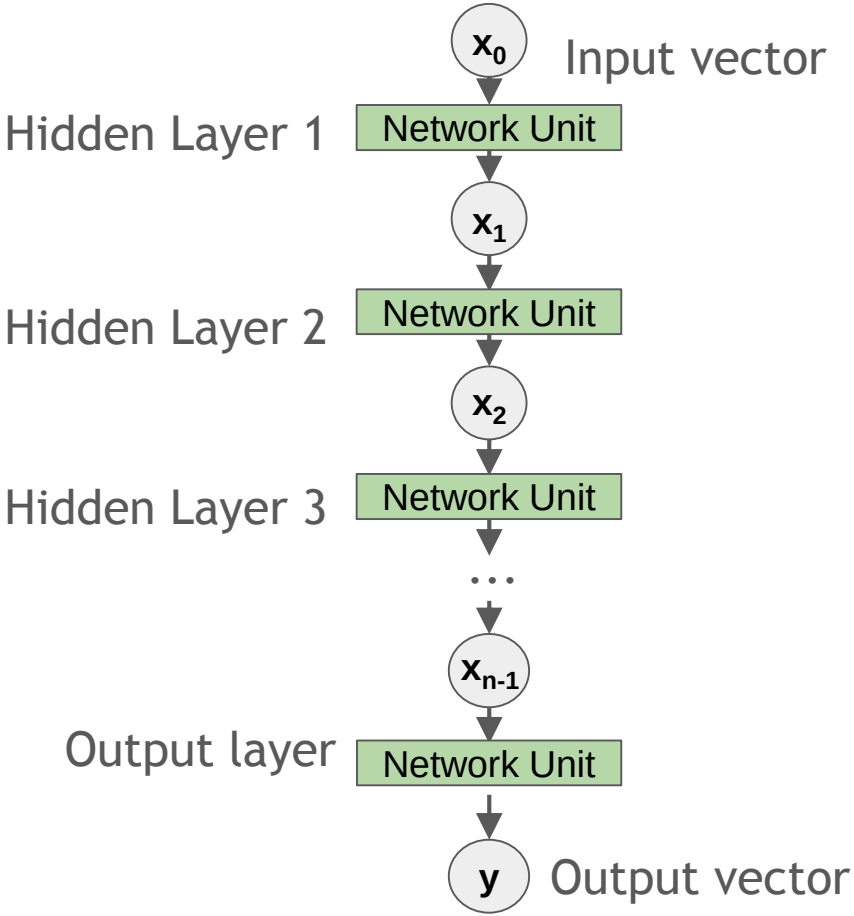


Testing error

- He et al.’s (2015) testing revealed that plain DNN suffered from higher training error, and as a result, a higher testing error.

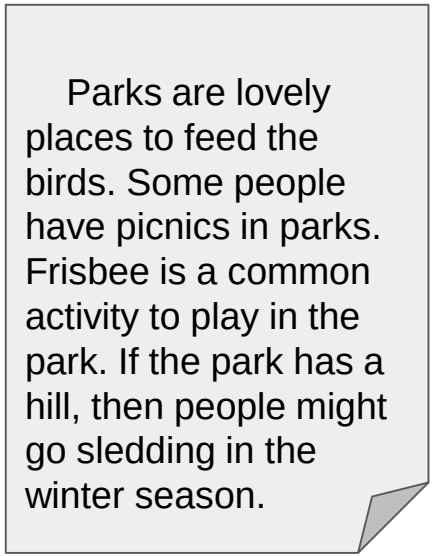
Residual Connections - The Difficulty of Training Deep Neural Networks

- Plain DNN are more difficult to train because:
  - Different parts of output lose accuracy
  - Difficult to achieve identity mappings
  - *Identity mapping* - output = input

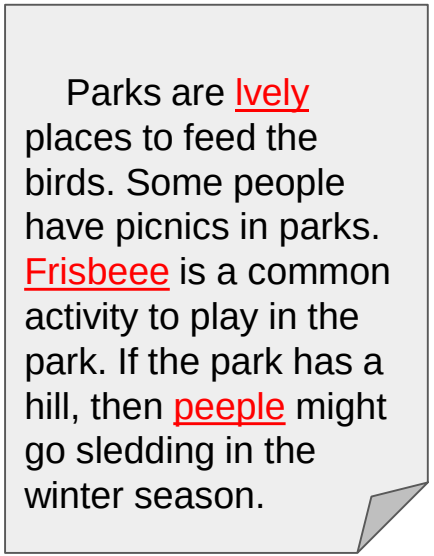


Residual Connections - Plain vs. Residual Mappings

- Perfectly retyping the original paragraph = identity mapping.



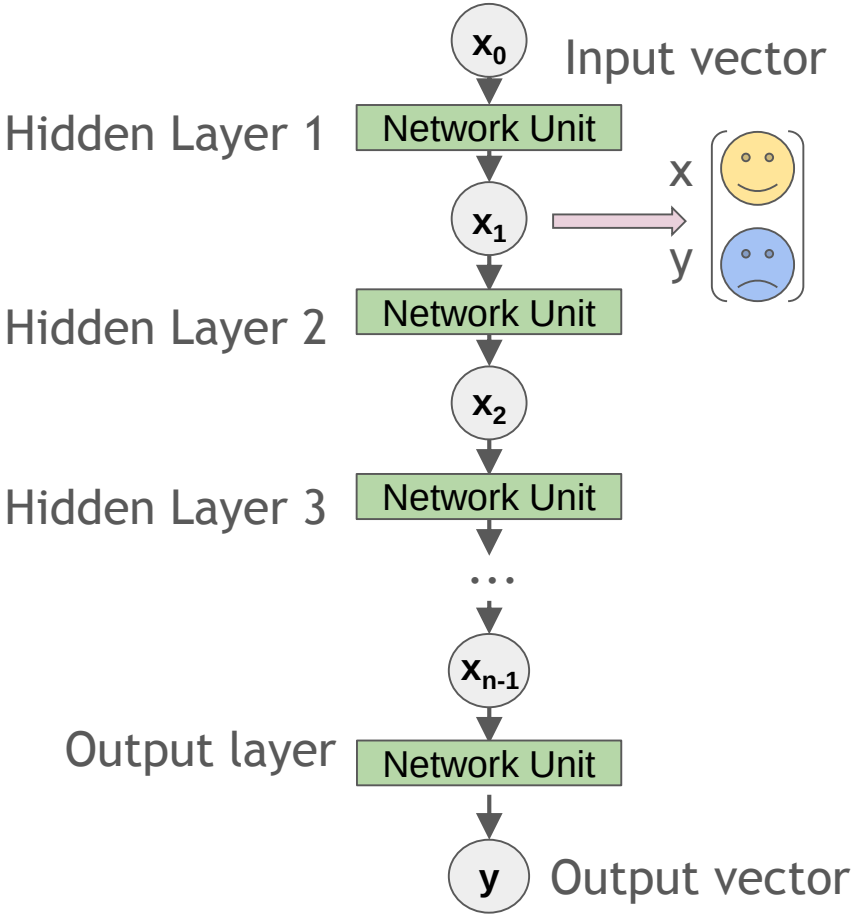
Original paragraph



Re-typed paragraph/
Plain Neural Network

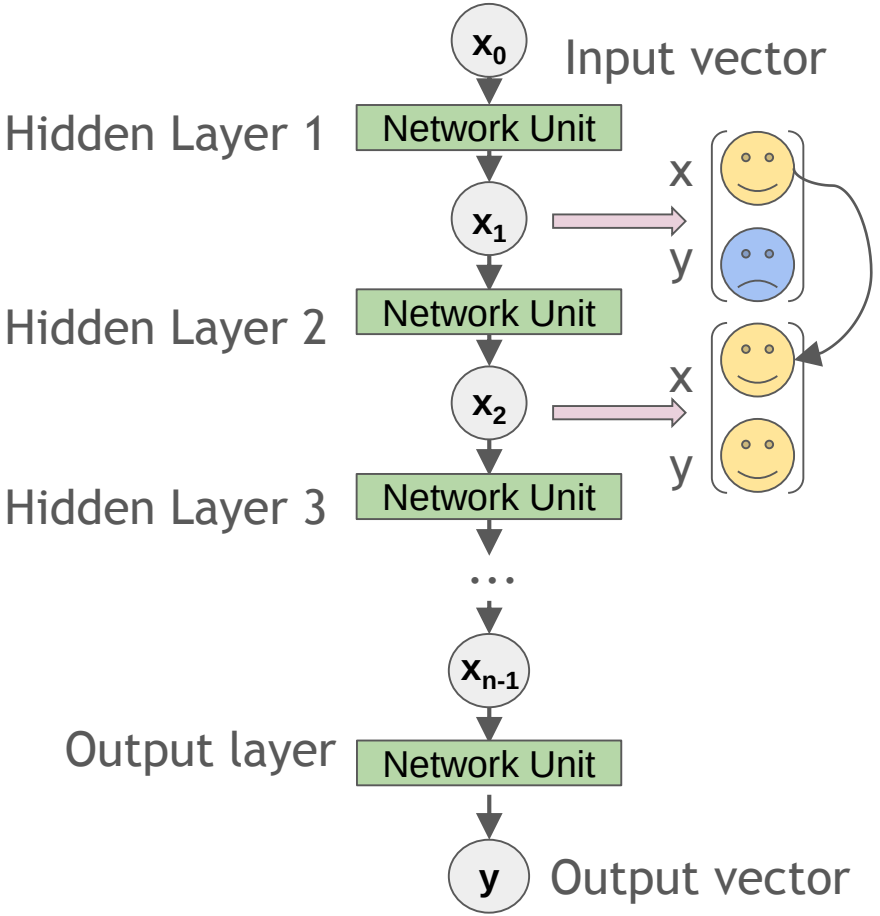
Residual Connections - The Difficulty of Training Deep Neural Networks

- Example:
  - x-axis value is perfect
  - y-value needs improvement



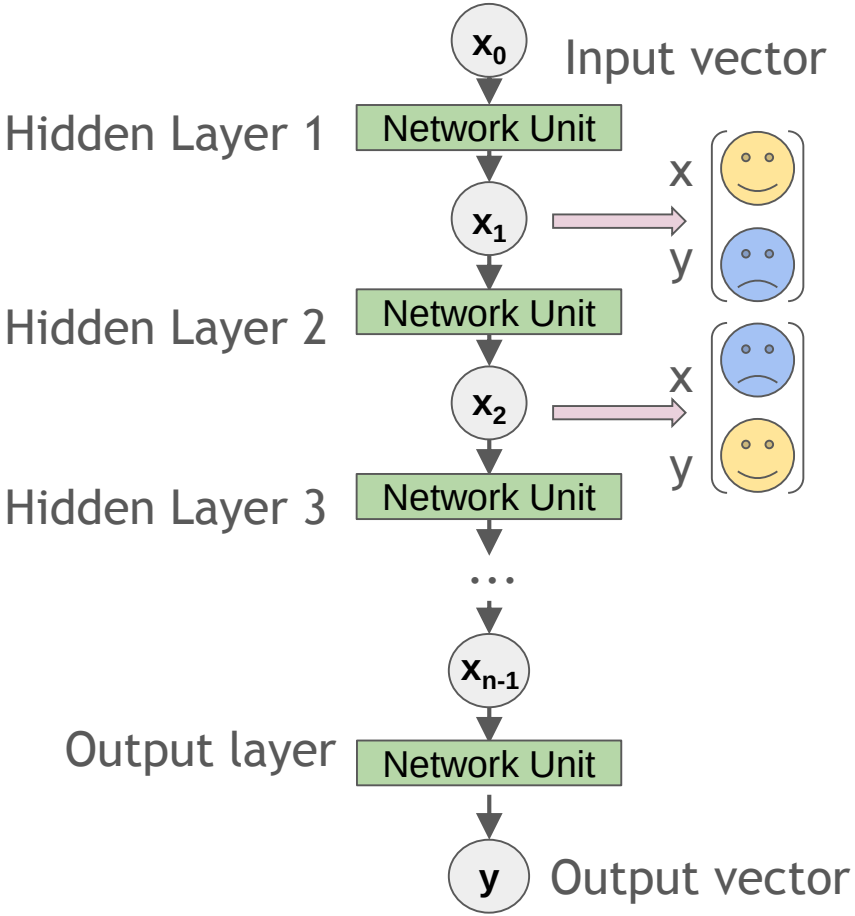
Residual Connections - The Difficulty of Training Deep Neural Networks

- Example:
  - x-axis value is perfect
  - y-value needs improvement
  - Goal: Maintain x, improve y



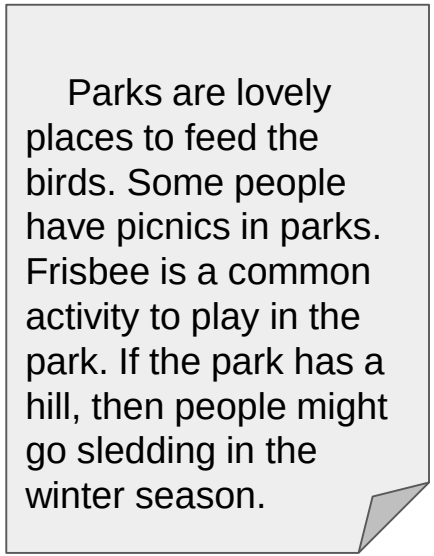
Residual Connections - The Difficulty of Training Deep Neural Networks

- Example:
  - x-axis value is perfect
  - y-value needs improvement
  - Goal: Maintain x, improve y
  - Reality: x degrades, y improves



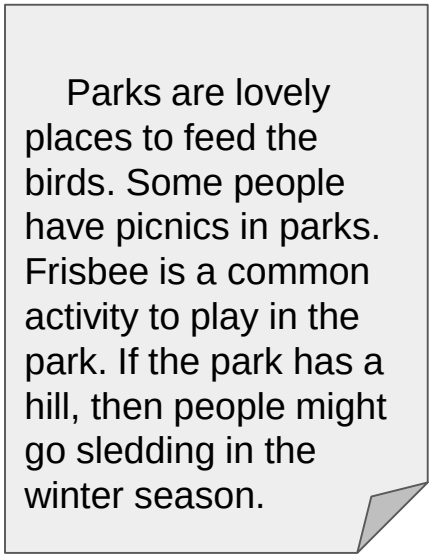
Residual Connections - Plain vs. Residual Mappings

- More room for error when re-typing a paragraph than when copy-and-pasting



Original paragraph

Re-typed paragraph/
Plain Neural Network

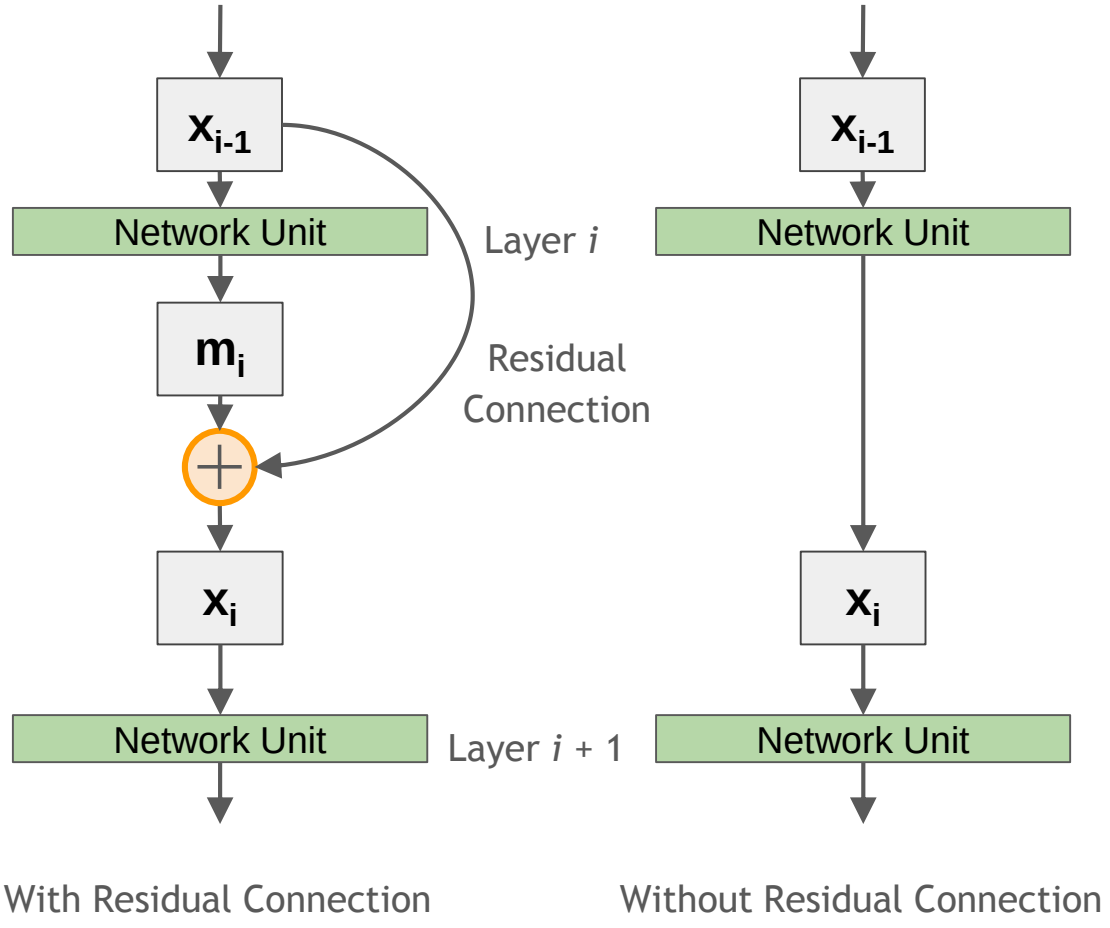


Copy-and-pasted paragraph/
Residually Connected Network

Residual Connections - Structure

Residual Connections

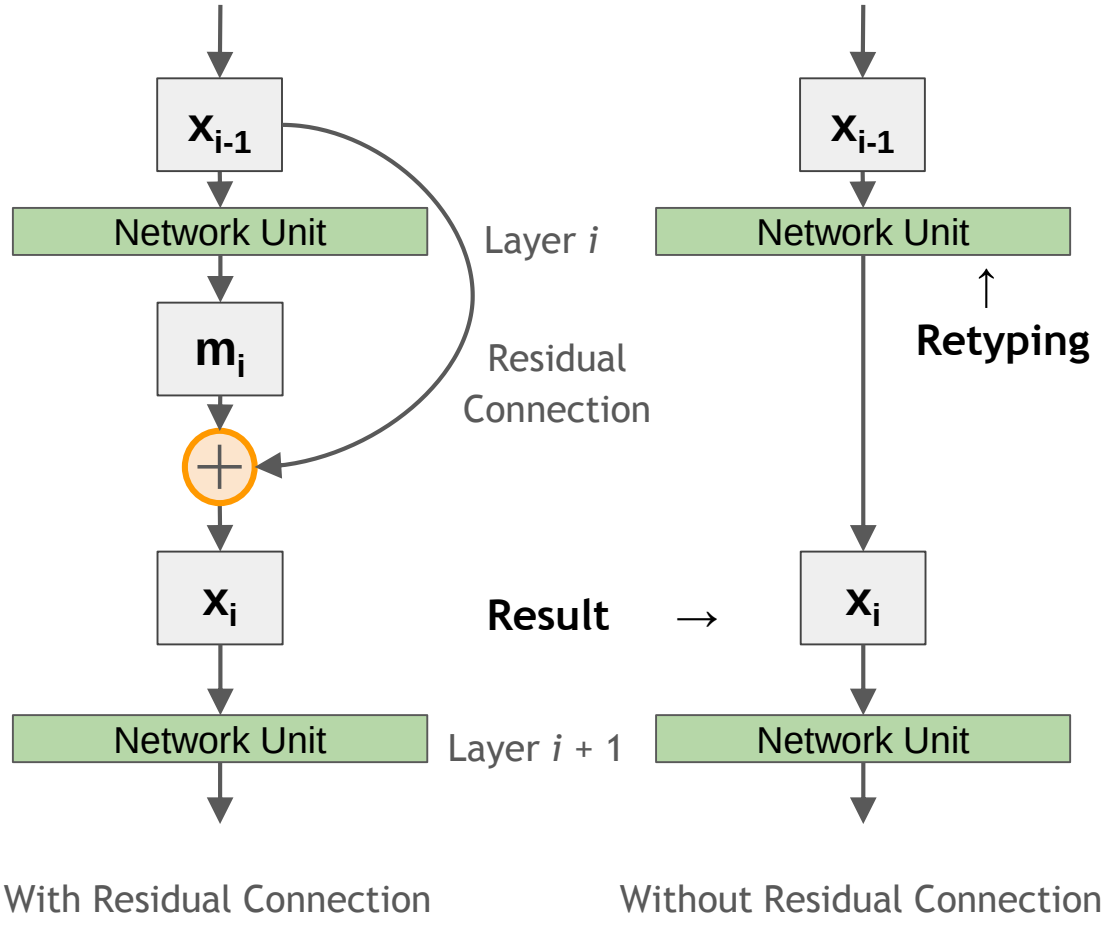
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ **when weights = 0**



Residual Connections - Structure

Residual Connections

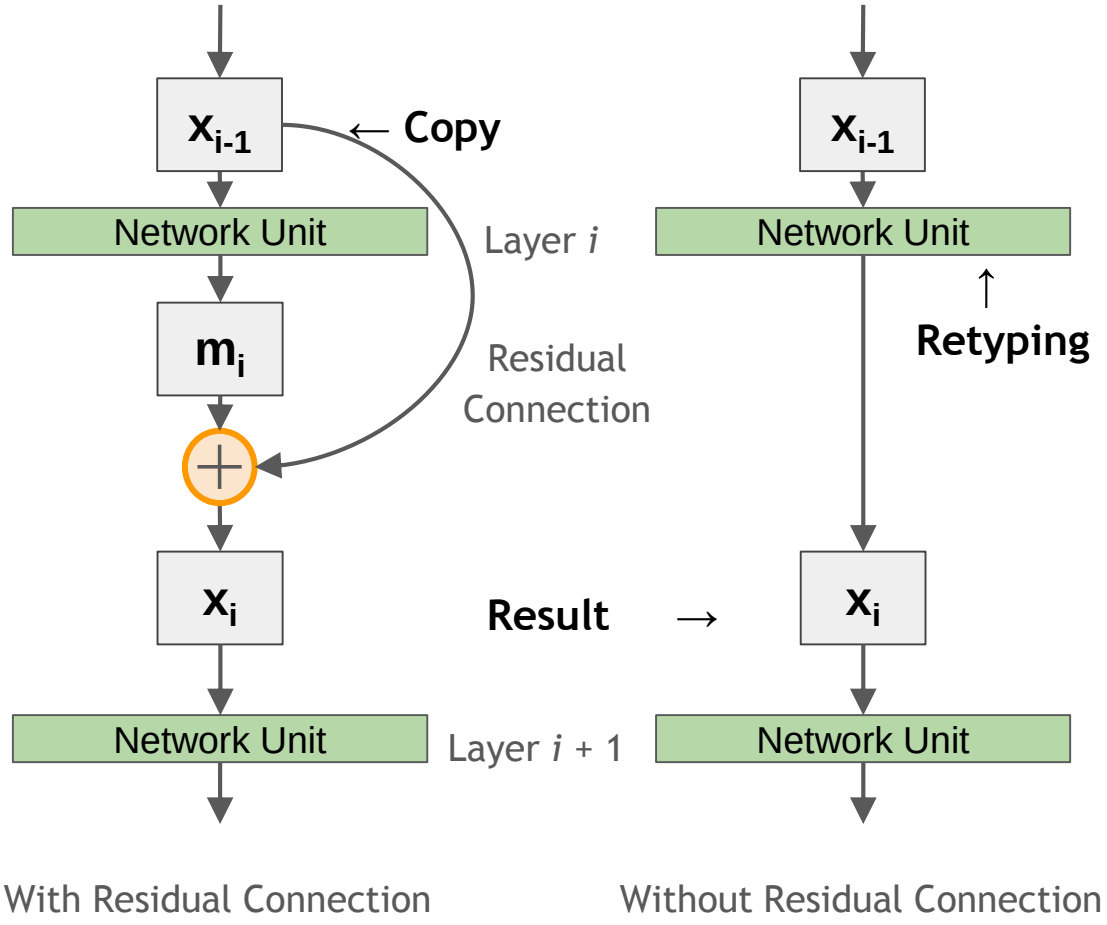
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ when weights = 0



Residual Connections - Structure

Residual Connections

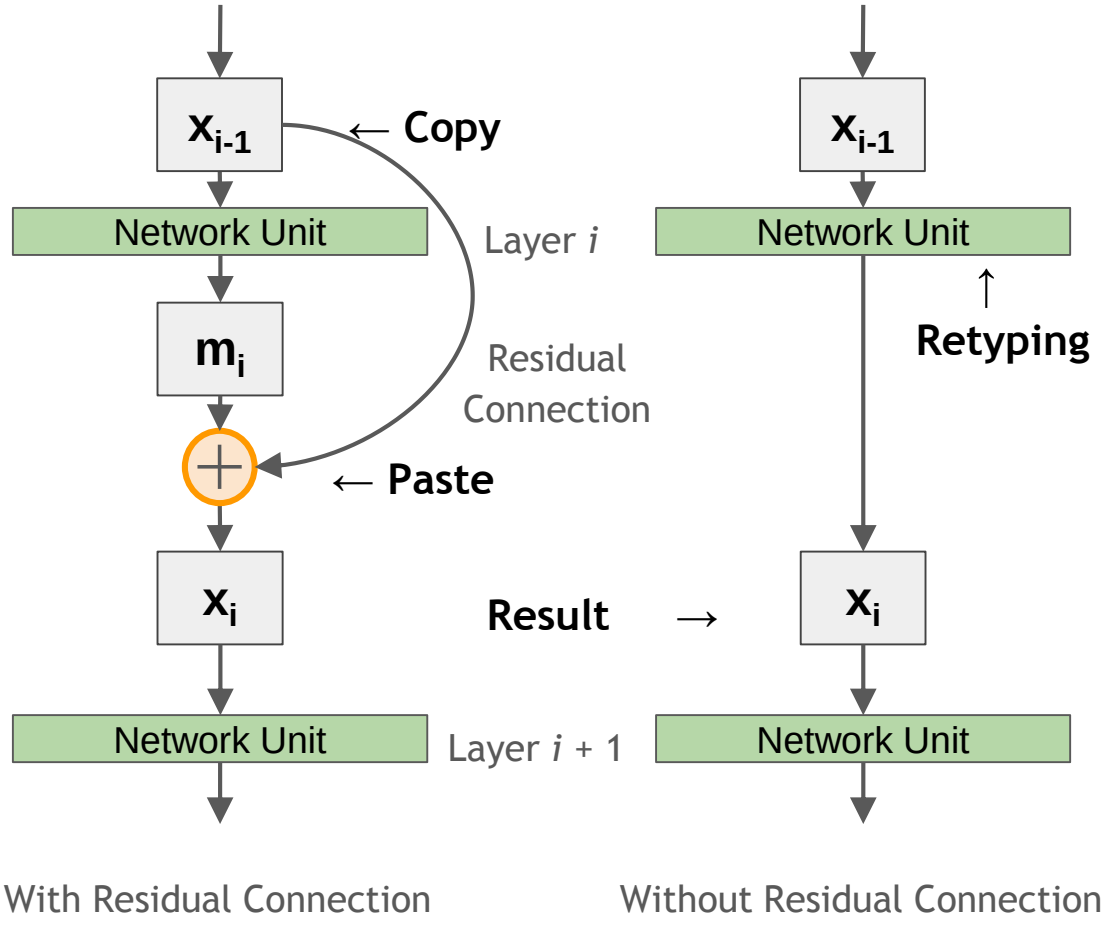
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ when weights = 0



Residual Connections - Structure

Residual Connections

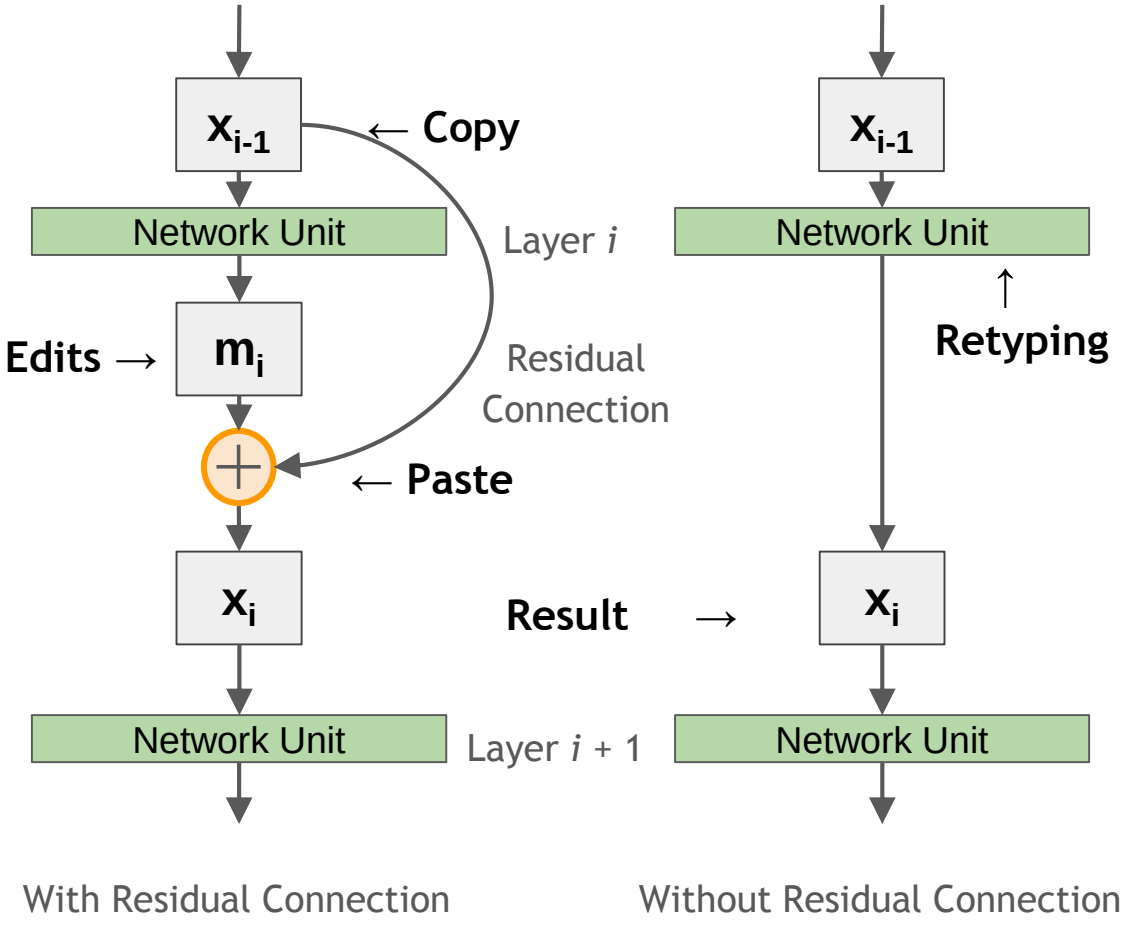
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ when weights = 0



Residual Connections - Structure

Residual Connections

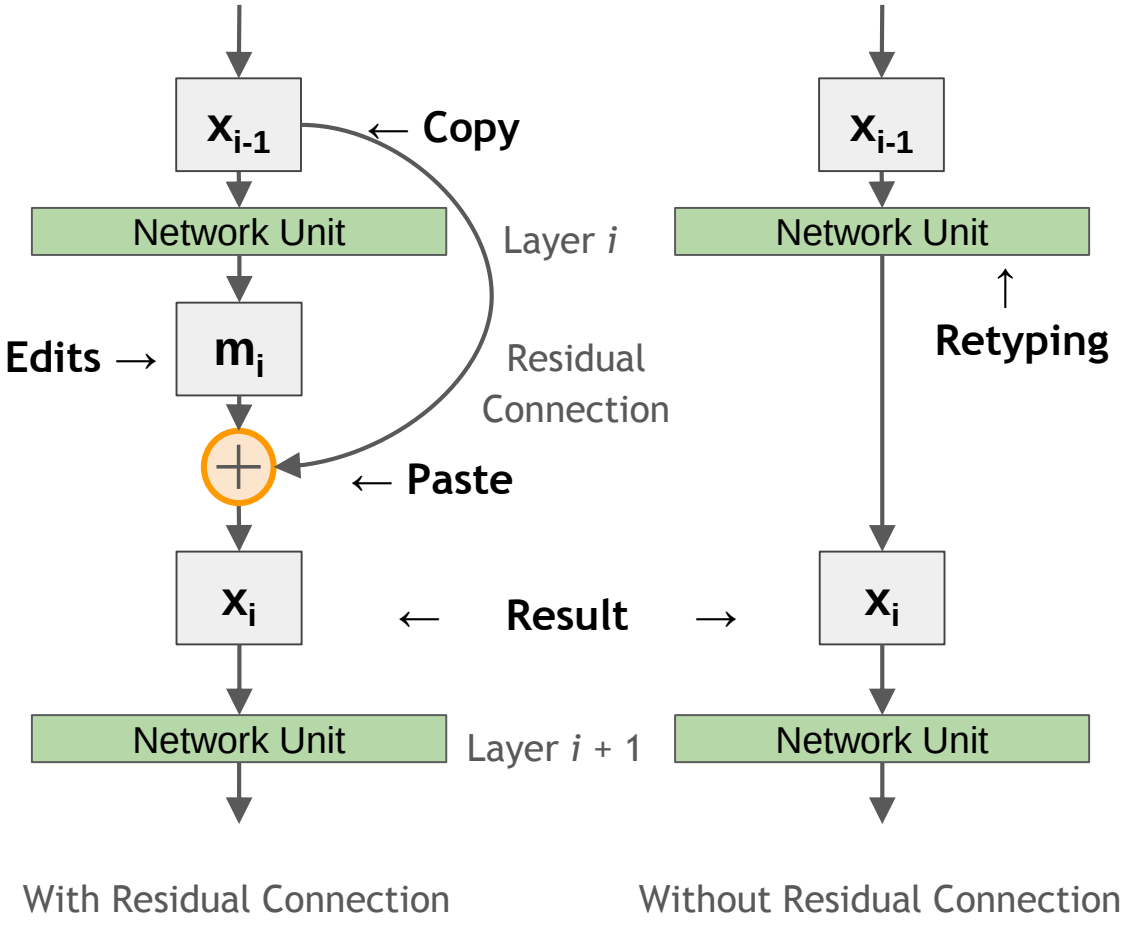
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ when weights = 0



Residual Connections - Structure

Residual Connections

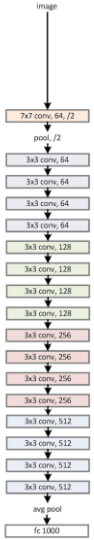
- layer *i* mapping: $x_i - x_{i-1}$
- $x_i = x_{i-1}$ when weights = 0



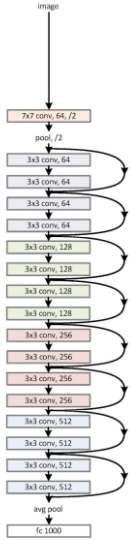
Residual Connections - Evaluations

- Google comparisons not available.
- He et al. (2015) compared four networks:
  - an 18-layer plain network
  - a 34-layer plain network
  - an 18-layer residual network
  - and a 34-layer residual network

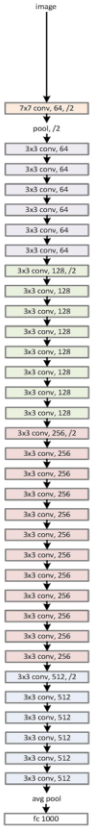
18-layer plain



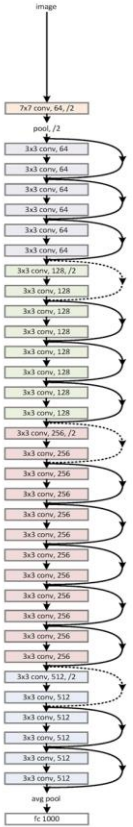
18-layer residual



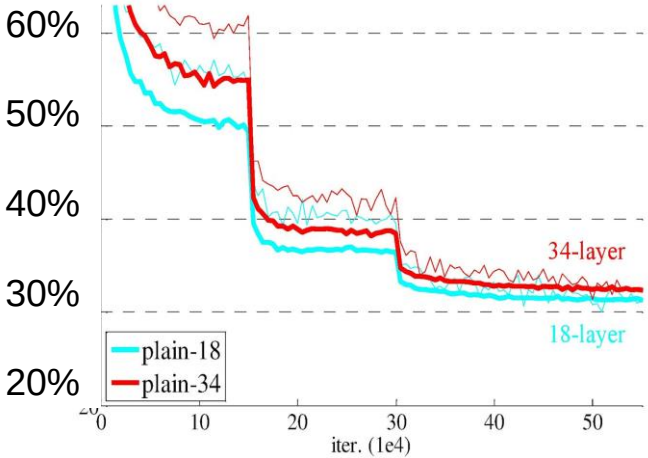
34-layer plain



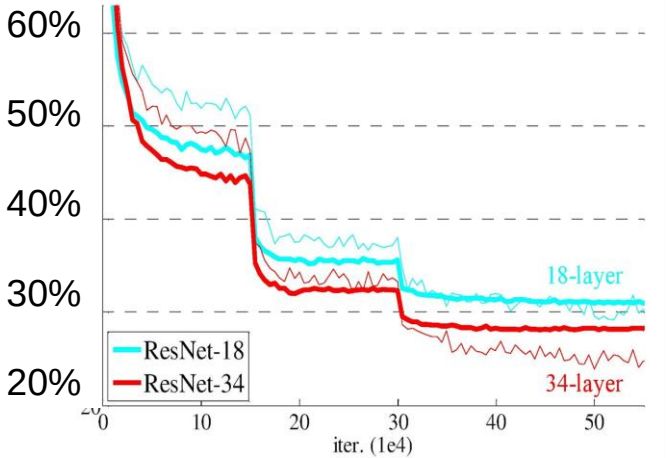
34-layer residual



Residual Connections - Evaluations



Plain



Residually Connected

Testing Error	plain	residual
18 layers	27.94%	27.88%
34 layers	28.54%	25.03%

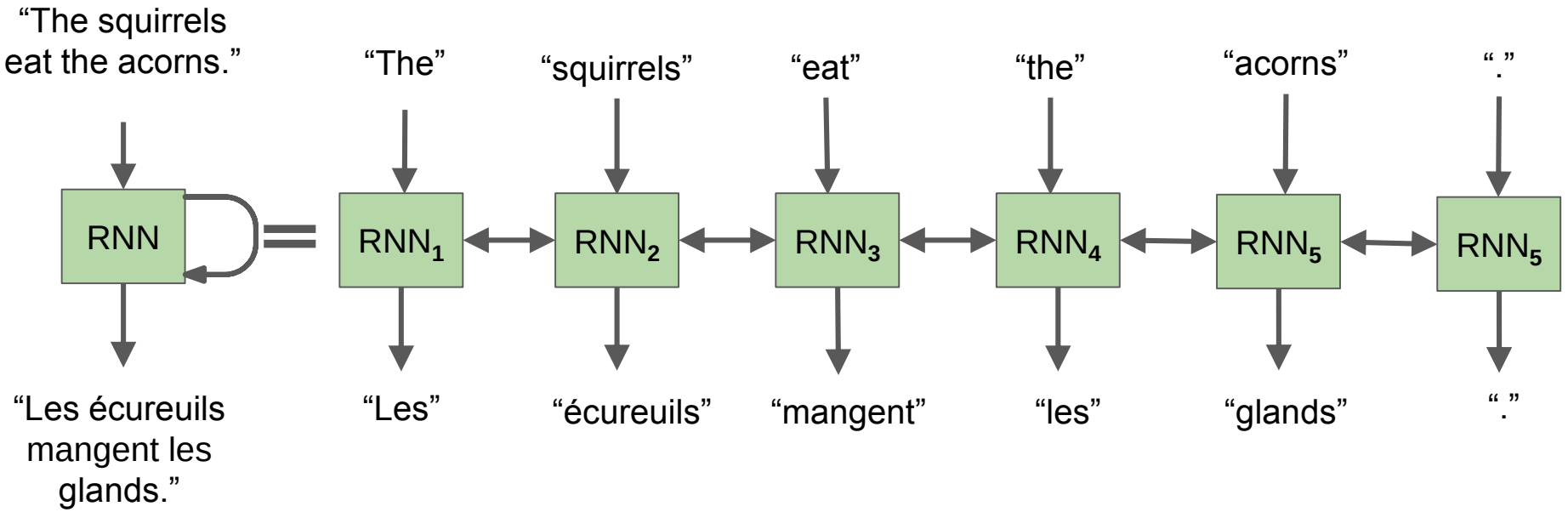
He et al. (2015)

Outline

- Introduction
- Neural Networks
- Residual Connections
- Attention Network
  - Recurrent Neural Networks
  - Encoder-Decoder Model
  - Attention Mechanism
  - Evaluations
- Summary

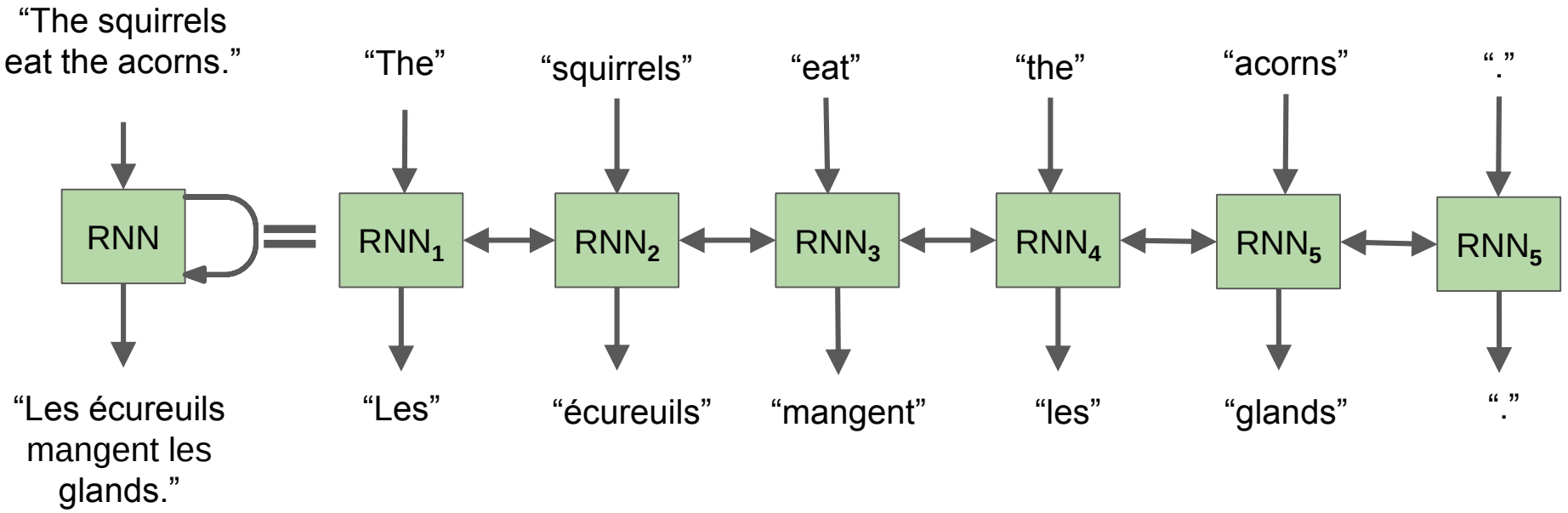
Attention Networks - Recurrent Neural Networks

- Recurrent Neural Networks (RNN) loop over the same unit multiple times.
  - Each iteration of the loop is known as a *time step*.



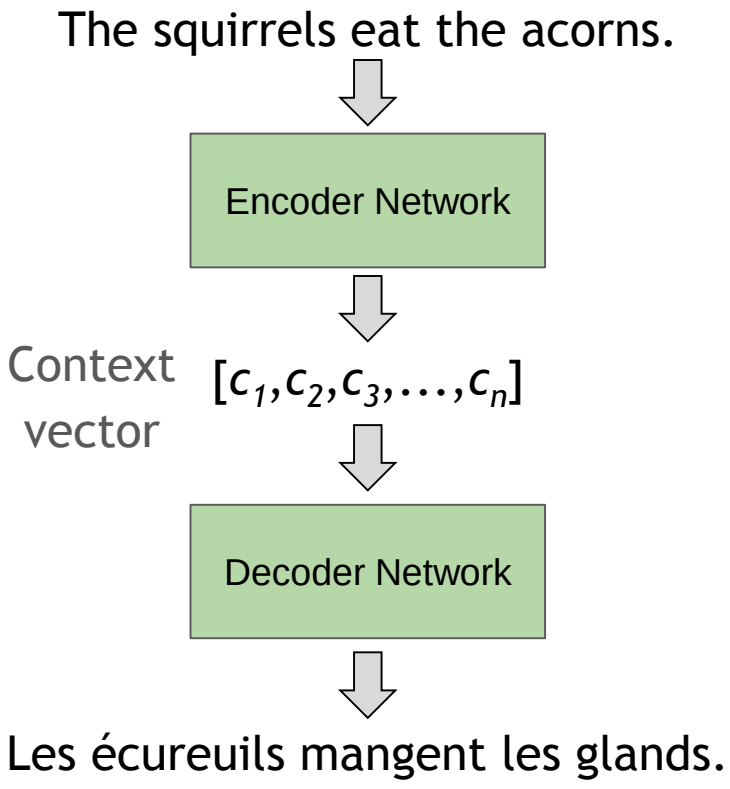
Attention Networks - Recurrent Neural Networks

- Bidirectional RNN - information travels forwards and backwards, giving words the context of words before and after them.



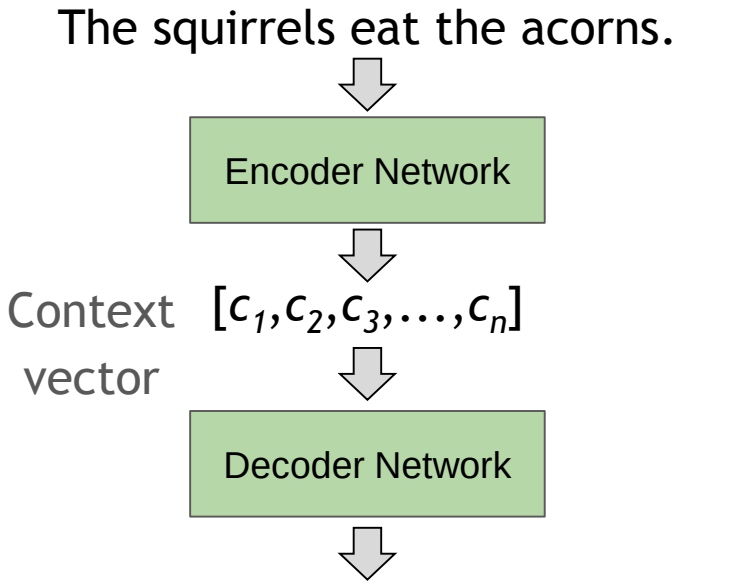
Attention Networks - Encoder-Decoder

- Two RNN
- Encoder network - produces context vector
- *context vector* - contains sentence information
- Decoder network - produces translated sentence



Attention Networks - Encoder-Decoder

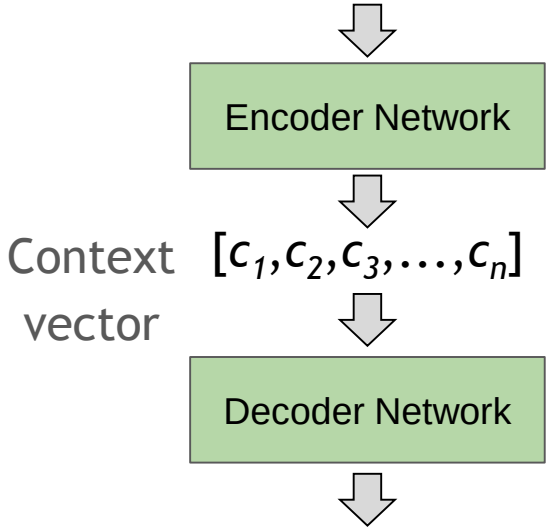
- Issues with longer sentences
- Context vector is a fixed size



Attention Networks - Encoder-Decoder

- Issues with longer sentences
- Context vector is a fixed size

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.



Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

Attention Networks - Translation Comparison

[Examples from Bahdanau et al. (2015)]

Input Sentence

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

Encoder-Decoder Translation

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

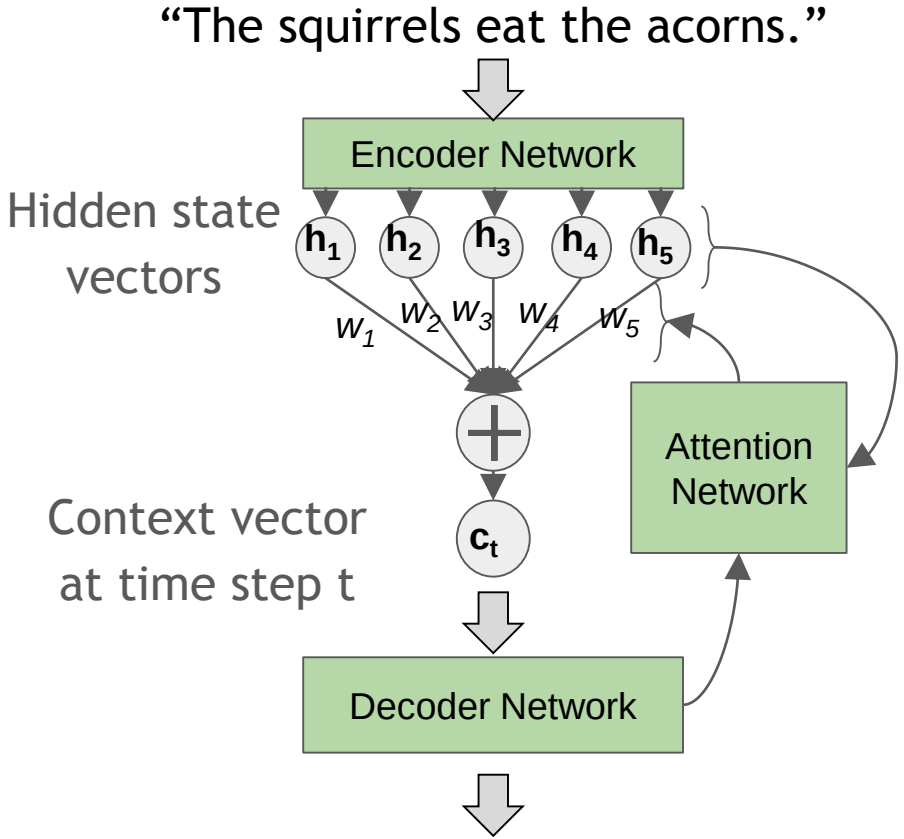
[based on his state of health]

Attention Network Translation

Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.

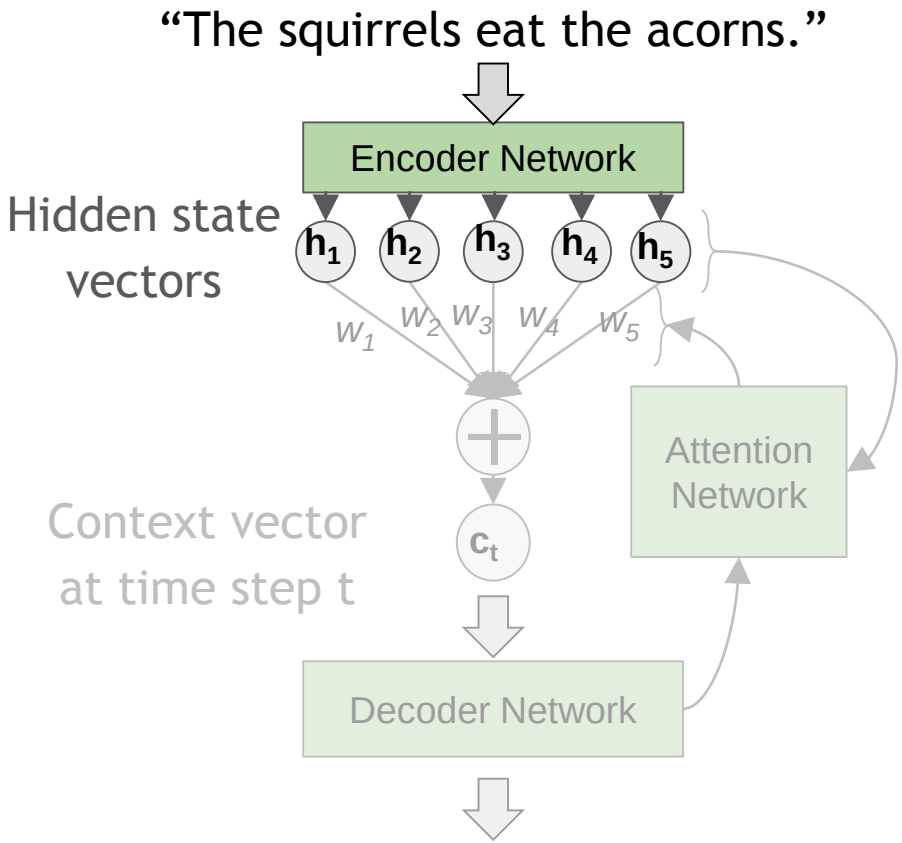
Attention Networks - Addition of an Attention Mechanism

- Attention network model extends encoder-decoder model.
- Three networks:
  - Encoder Network
  - Attention Network
  - Decoder Network



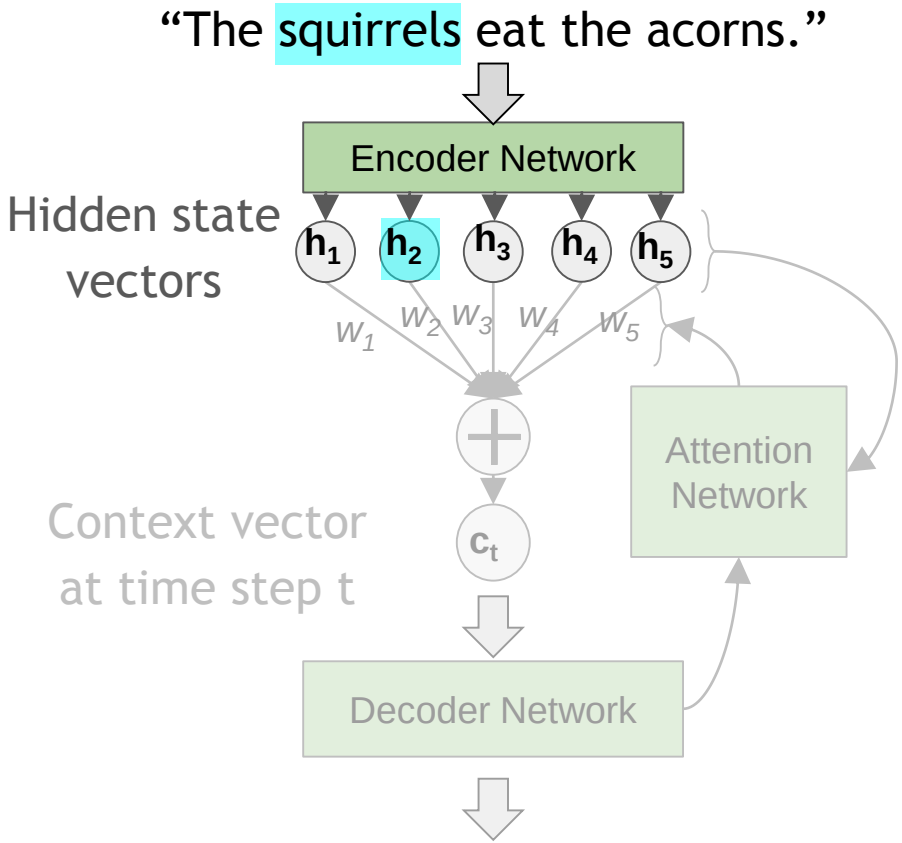
Attention Networks - Addition of an Attention Mechanism

- Encoder produces *hidden state vectors*.
- Hidden state vectors
  - One for each input word
  - Includes information about the whole input sentence
  - Focuses strongly on what surrounds its word



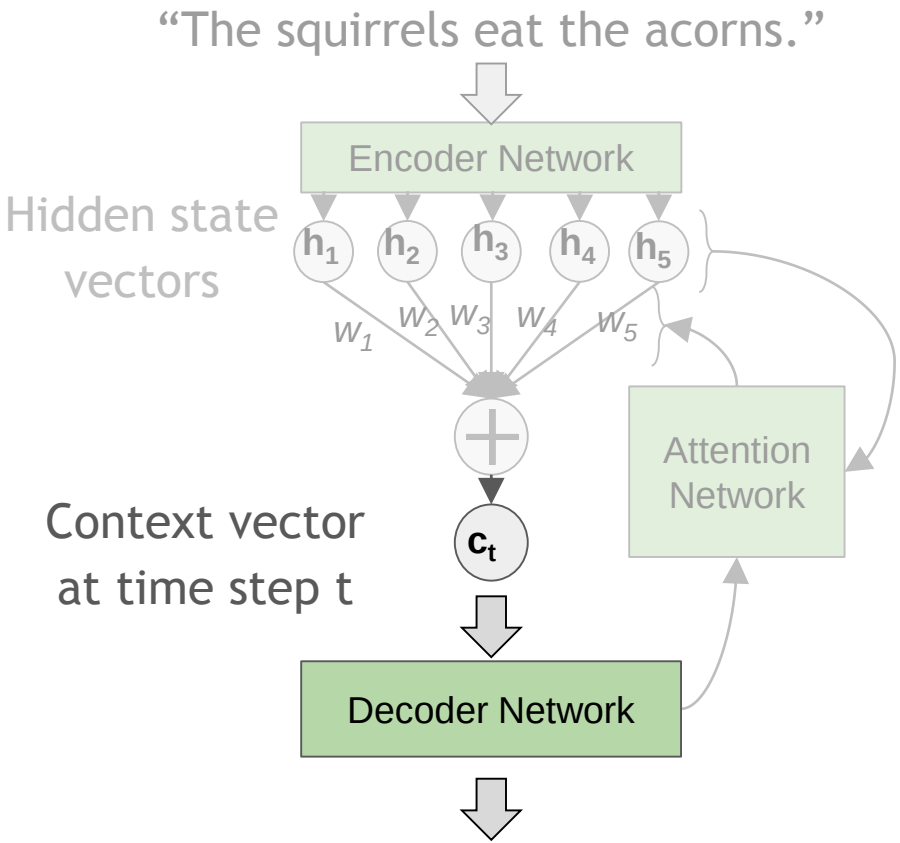
Attention Networks - Addition of an Attention Mechanism

- Encoder produces *hidden state vectors*.
- Hidden state vectors
  - One for each input word
  - Includes information about the whole input sentence
  - Focuses strongly on what surrounds its word



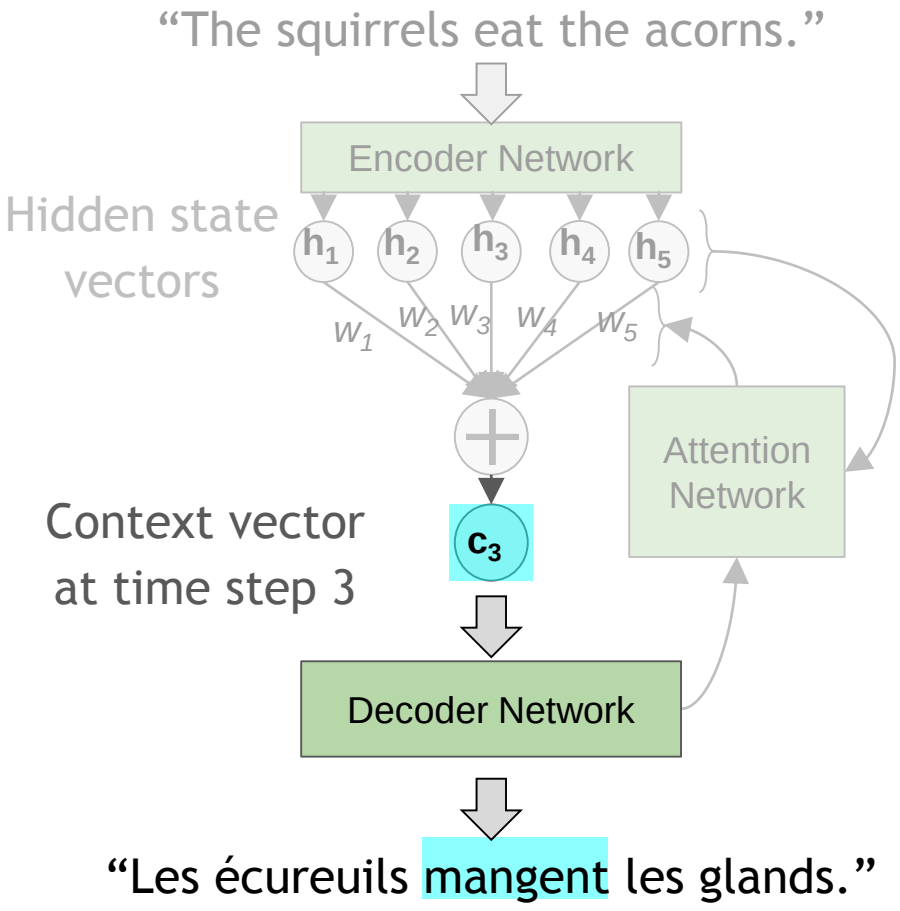
Attention Networks - Addition of an Attention Mechanism

- Context vector for each output word
- Decoder uses t^{th} context vector to translate t^{th} word



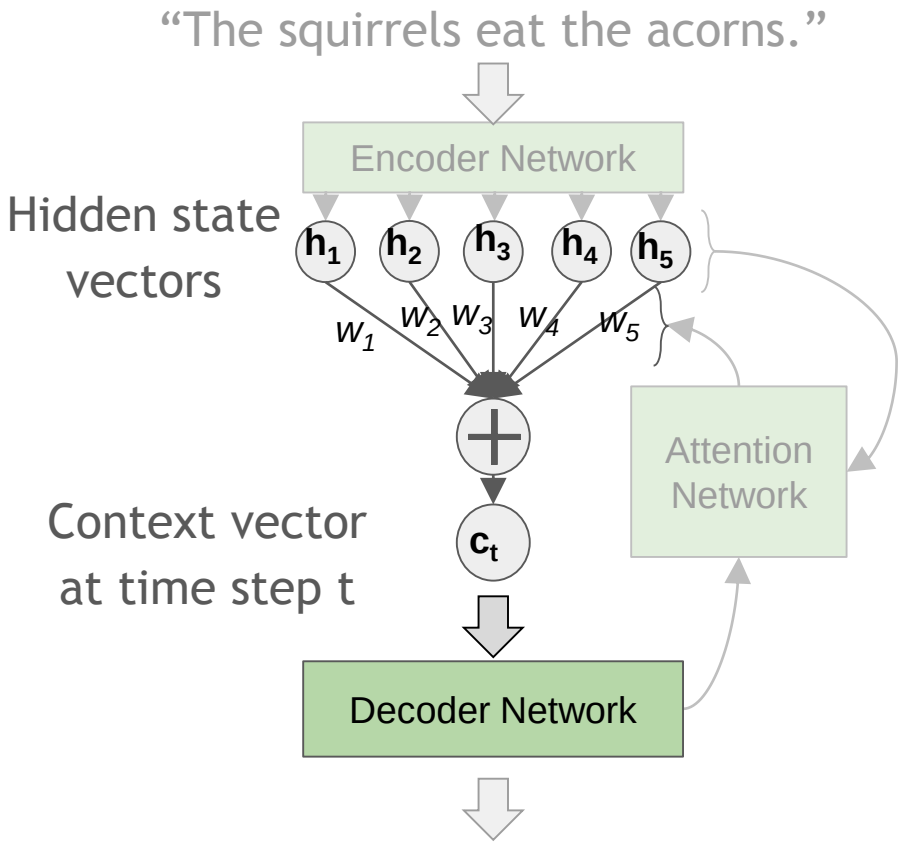
Attention Networks - Addition of an Attention Mechanism

- Context vector for each output word
- Decoder uses t^{th} context vector to translate t^{th} word

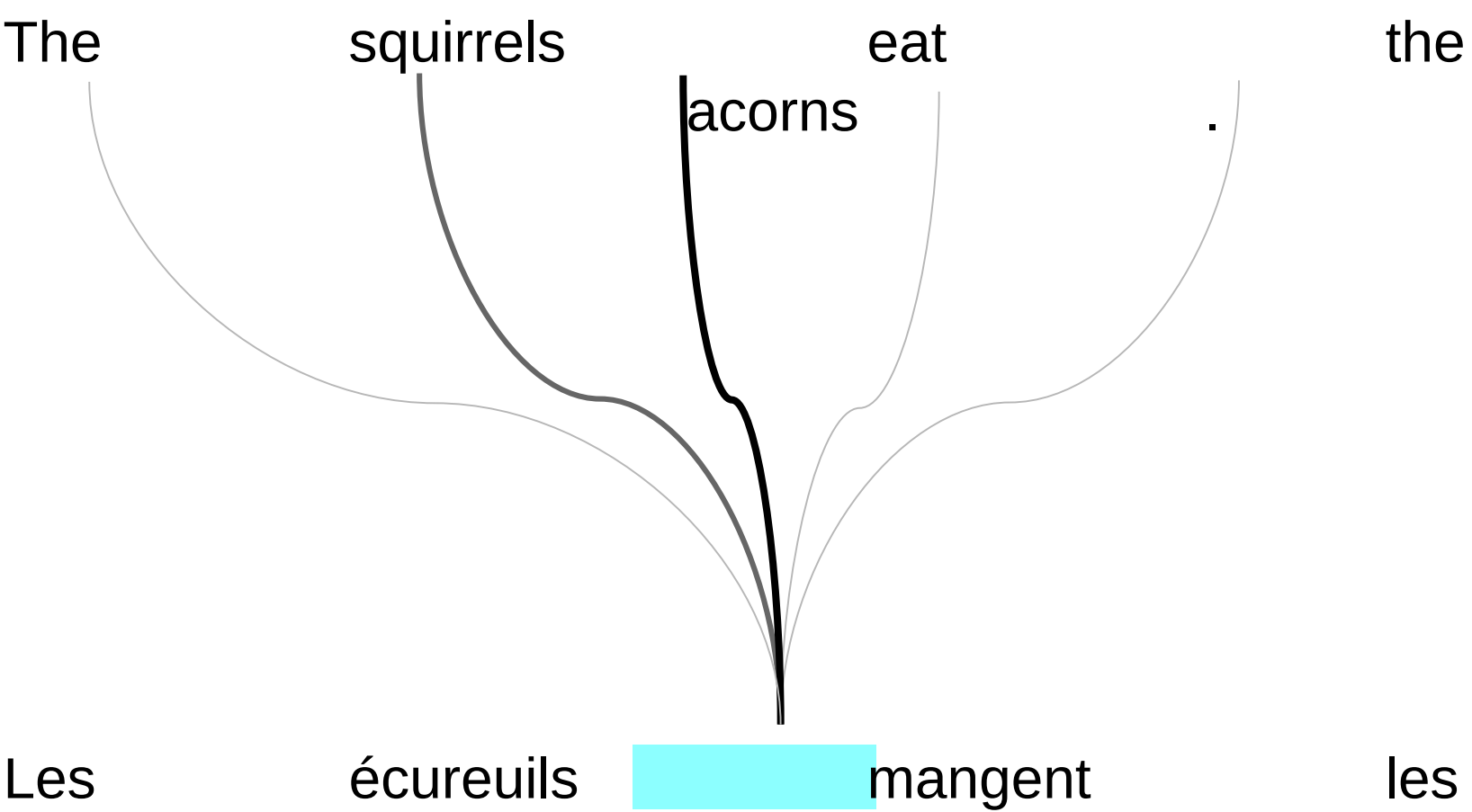


Attention Networks - Addition of an Attention Mechanism

- t^{th} context vector = sum of the weighted hidden state vectors.
- Weights are recalculated for each time step.
  - determine strength of input word's effect on output word.

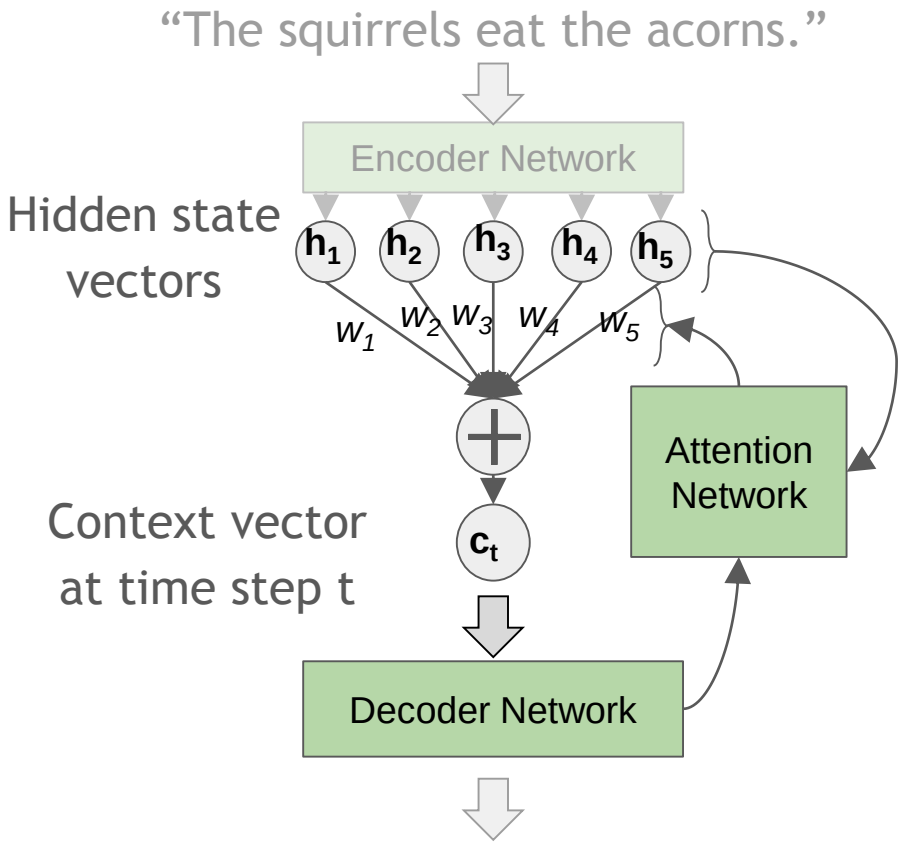


Attention Mechanisms - Word Influence



Attention Networks - Addition of an Attention Mechanism

- *Attention mechanism*
  - Gives decoder more information for longer sentences, and less information for shorter sentences
- Attention Network
  - Determines the weights of the hidden state vectors
  - Takes in information from decoder and hidden state vectors.

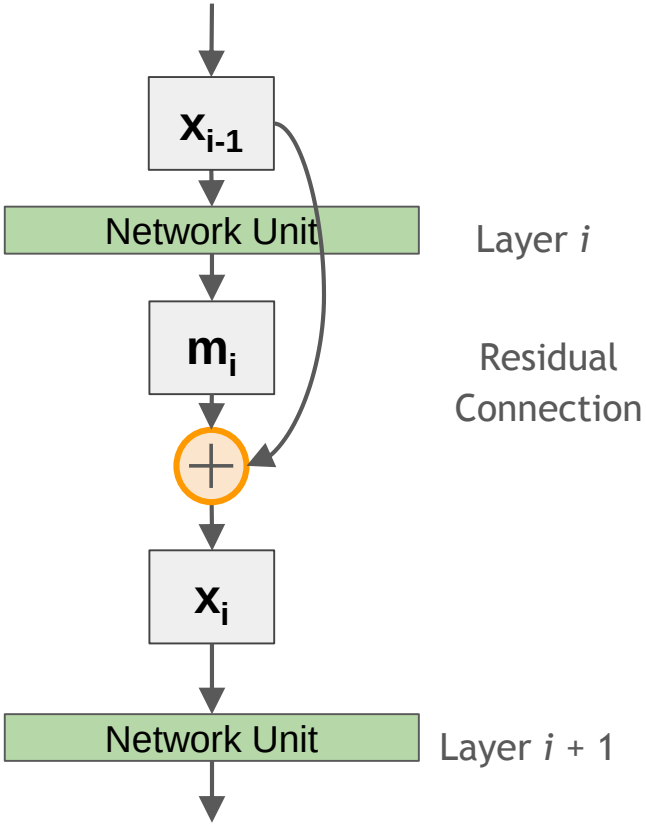


Attention Networks - Evaluations

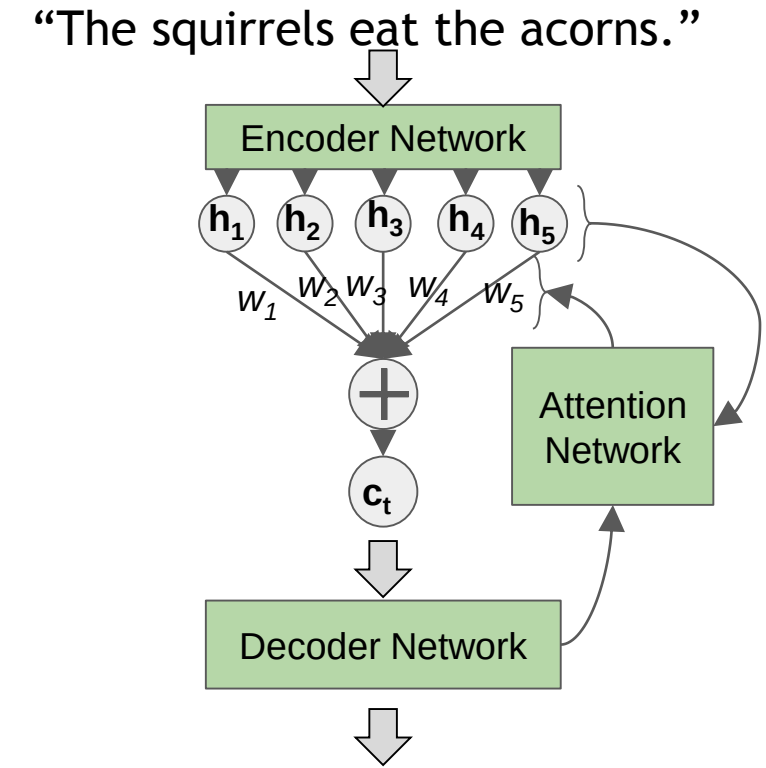
- Google comparisons not available.
- Bahdanau et al. (2015) evaluated an encoder-decoder model against an attention network model using the ACL WMT ‘14 dataset.
- Higher BLEU scores mean the translation is closer to human translation.

	Model	BLEU Score
Encoder-Decoder Model	RNNencdec-30	13.93
	RNNencdec-50	17.82
Attention Network Model	RNNsearch-30	21.50
	RNNsearch-50	26.75

Summary



Residual Connections



Attention Network Model

Acknowledgements

Thank you for your time and attention!

Thank you to my advisor Elena Machkasova, KK Lamberty and Mitchell Finzel for your guidance and feedback.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural Machine Translation by jointly learning to align and translate.IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition.IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi,W. Macherey, M. Krikun, Y. Cao, Q. Gao,K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo,H. Kazawa, K. Stevens, G. Kurian, N. Patil,W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick,O. Vinyals, G. Corrado, M. Hughes, and J. Dean.Google's neural machine translation system: Bridging the gap between human and Machine Translation. arXiv preprint arXiv:1609.08144, 2016.

[4] M. Sundermeyer, H. Ney, and R. Schluter. From feedforward to recurrent lstm neural networks for language modeling.IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 3, MARCH 2015, 2015.

Questions?

Attention Mechanisms - Word Influence

English Sentence:

The girls played tag and then they became tired.

Incorrect French Translation:

Les filles ont joué une étiquette et ensuite ils sont devenus fatigués.

Correct French Translation:

Les filles ont joué une étiquette et ensuite elles sont devenues fatiguées.