AMTA 2018
March 17 - 21, 2018
Boston, MA, USA

The 13th Conference of
The Association for Machine Translation
in the Americas

www.conference.amtaweb.org

TUTORIAL
March 17, 2018

De-mystifying Neural MT

Presenters: Dragos Munteanu *(SDL)*, Ling Tsou *(SDL)*

De-mystifying Neural MT

Dragos Munteanu

Ling Tsou

What you will get out of this tutorial

- Learn what’s behind the “magic”
- Make sense of the “buzzwords”
- Gain insights about why Neural Networks are so successful
- Better understand the limitations/difficulties in this new paradigm

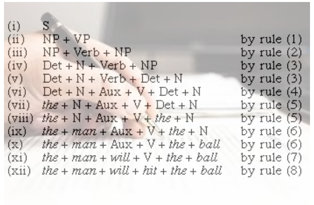
Who are we

- Dragos Munteanu
  - Director of Research and Development
  - 10+ years of experience
  - Started out at Language Weaver
- Ling Tsou
  - Research Engineer
  - 5+ years of experience

Agenda

- Neural Networks
  - Basic structure of a Neural Network
  - Deep Neural Networks
  - Training
- Neural Machine Translation
  - NMT vs SMT
  - Word embeddings
  - Architectures
  - Limitations
  - Future Outlook

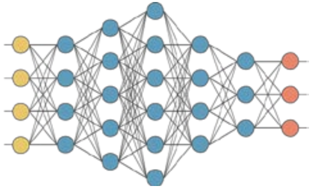
Rule-based vs. Statistical vs. Neural



Rule-Based

$$\tilde{e} = arg \max_{e \in e^*} p(e|f)$$

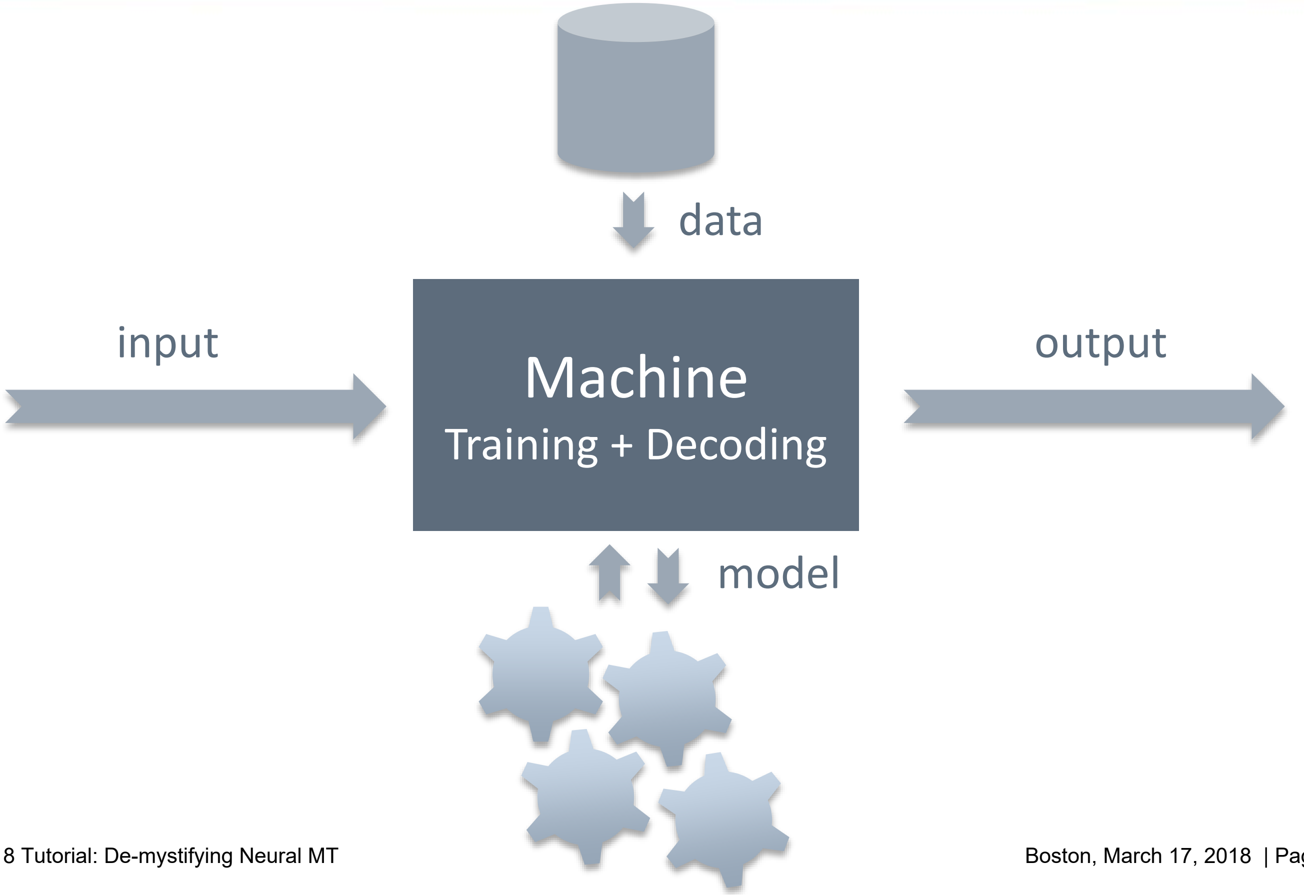
Statistical



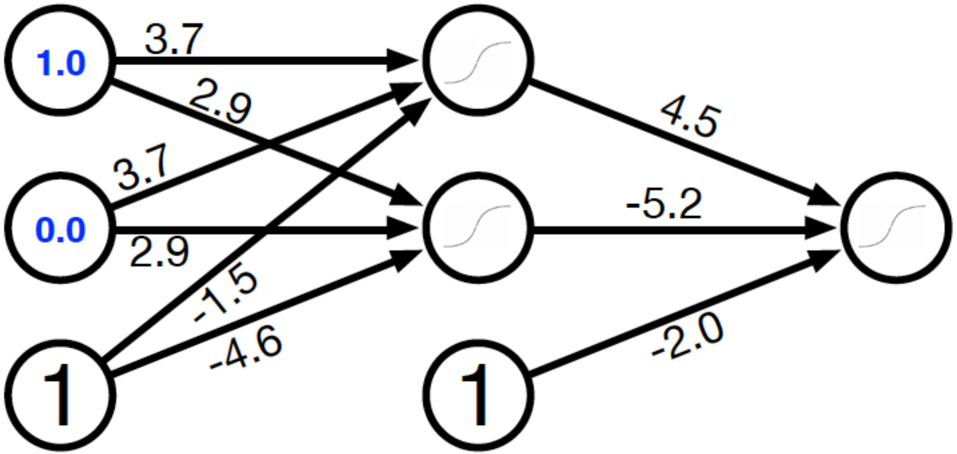
Neural


FIAT

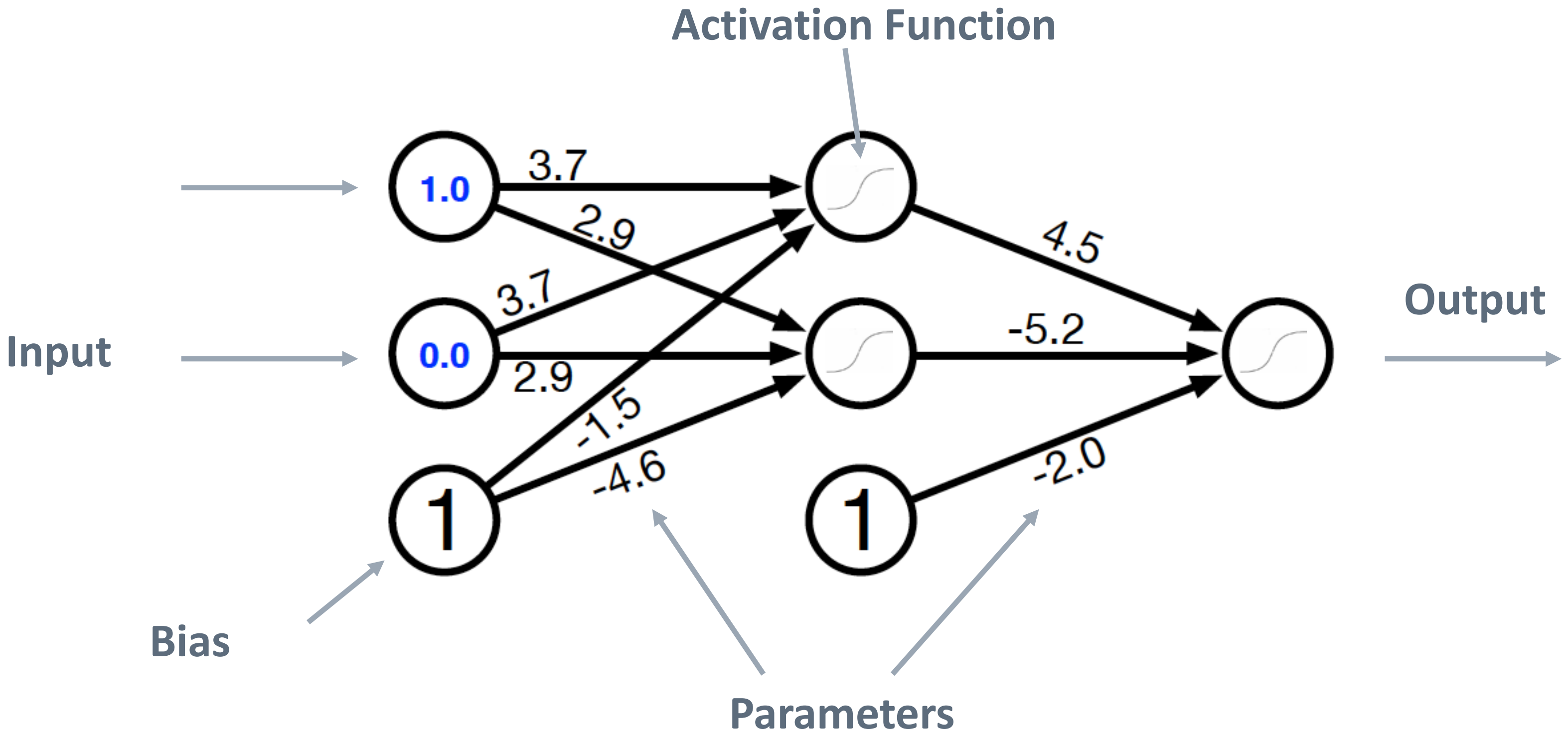
Statistical Learning



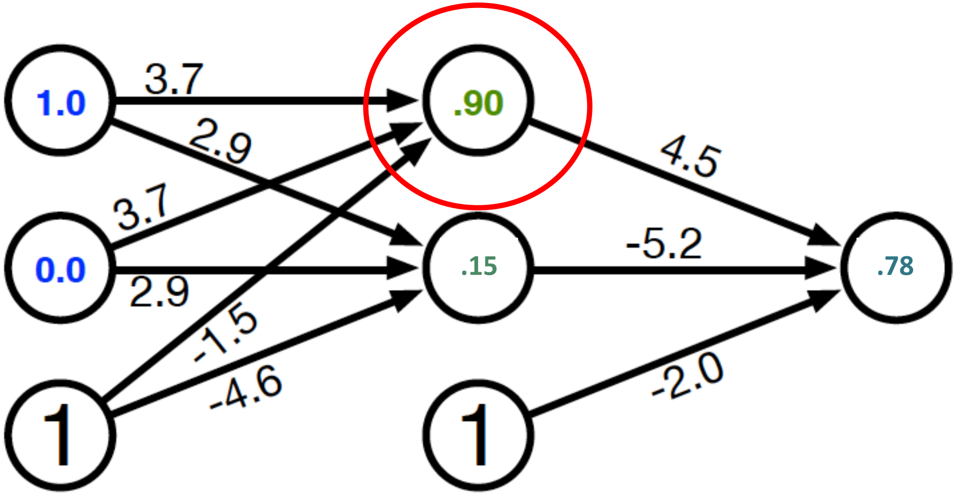
A Neural Network



A Neural Network

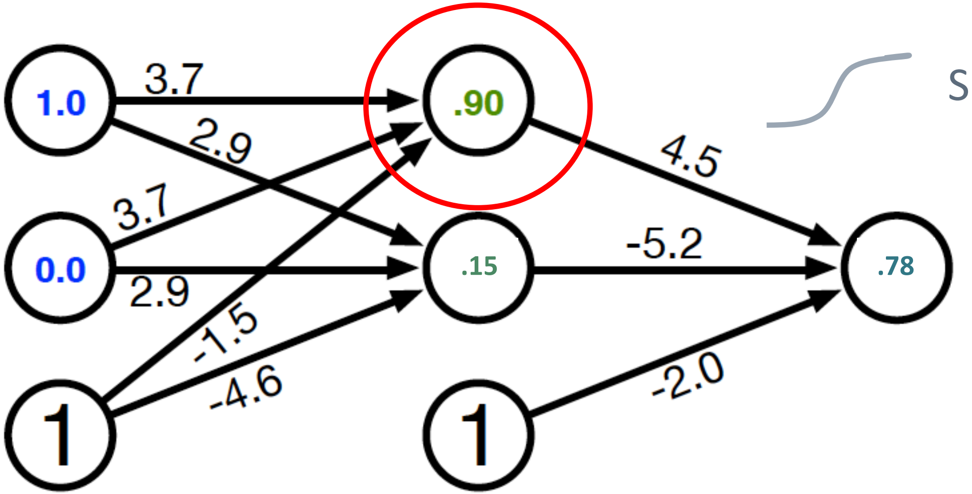


A Neural Network

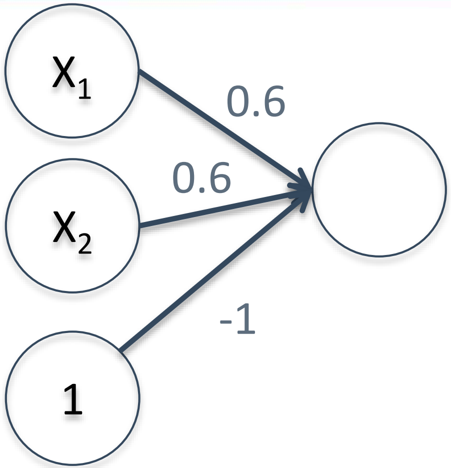


A Neural Network

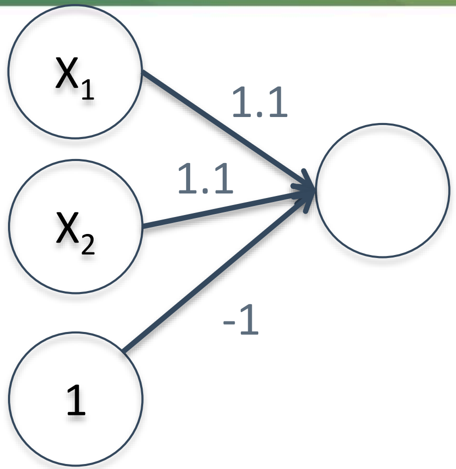
$1.0 \times 3.7 + 0.0 \times 3.7 + 1 \times -1.5 = 2.2$ $\rightarrow$ $\frac{1}{1+e^{-2.2}} = 0.90$



Sigmoid $f(x) = \frac{1}{1+e^{-x}}$



X_1	X_2	Target	*Network output*
0	0	0	*-1*
0	1	0	*-0.4*
1	0	0	*-0.4*
1	1	1	*0.2*

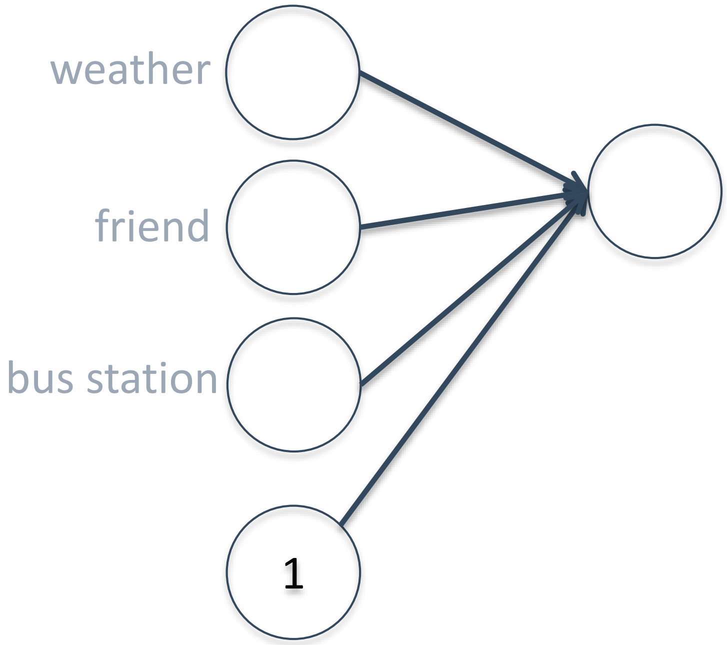


X_1	X_2	Target	*Network output*
0	0	0	*-1*
0	1	1	*0.1*
1	0	1	*0.1*
1	1	1	*0.2*

Should you go to the cheese festival?

Decision factors:

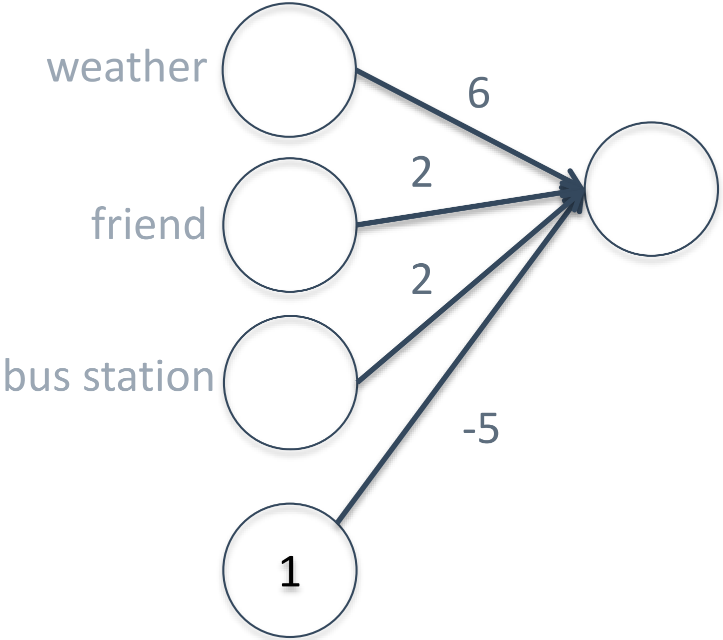
- Is weather good?
- Is friend coming?
- Is festival near bus station?



Should you go to the cheese festival?

Decision factors:

- Is weather good?
- Is friend coming?
- Is festival near bus station?

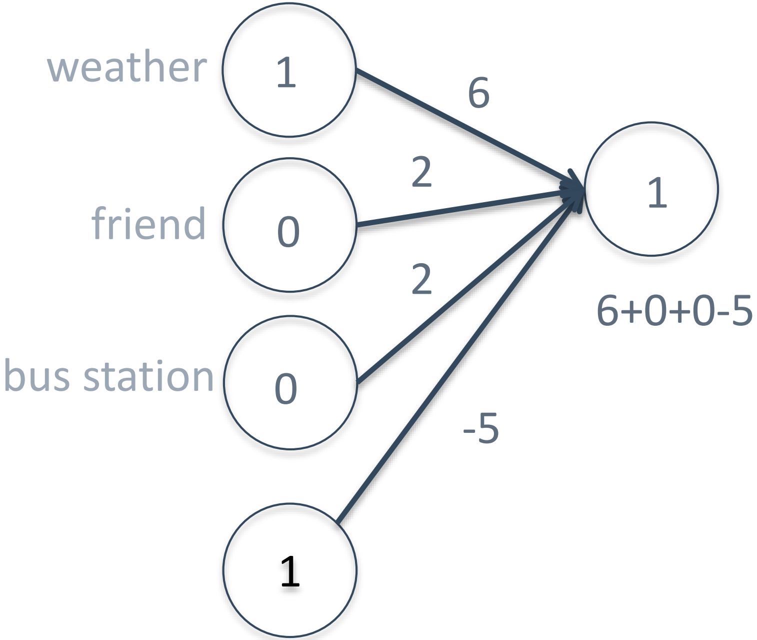


Going, unless weather is bad

Should you go to the cheese festival?

Decision factors:

- Is weather good?
- Is friend coming?
- Is festival near bus station?

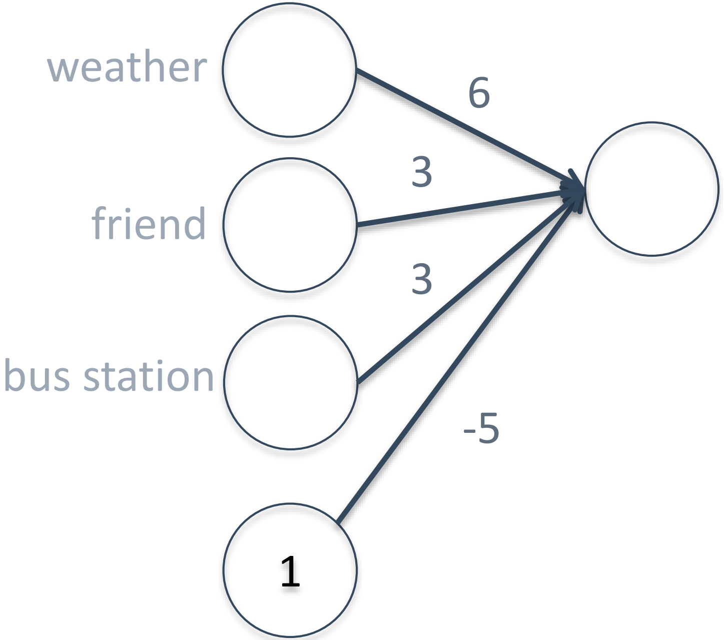


Going, unless weather is bad

Should you go to the cheese festival?

Decision factors:

- Is weather good?
- Is friend coming?
- Is festival near bus station?

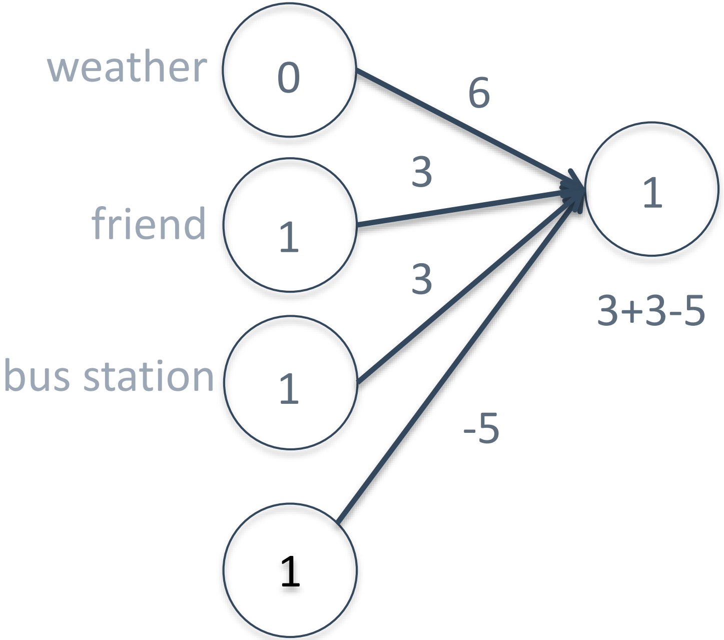


Going if weather is good OR friend+bus

Should you go to the cheese festival?

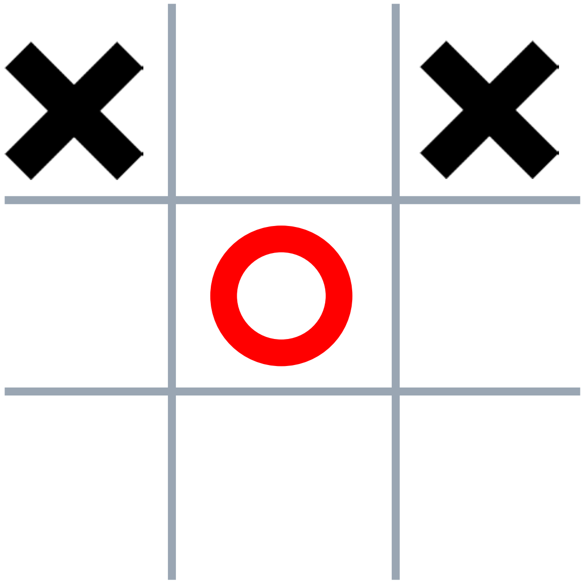
Decision factors:

- Is weather good?
- Is friend coming?
- Is festival near bus station?



Going if weather is good OR friend+bus

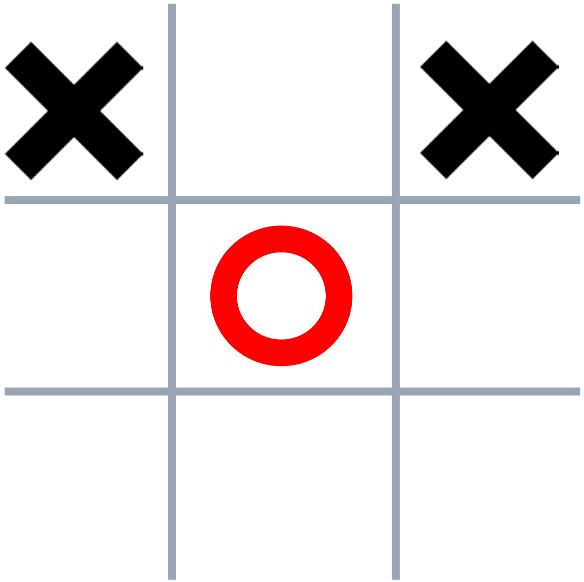
Playing games: Tic Tac Toe

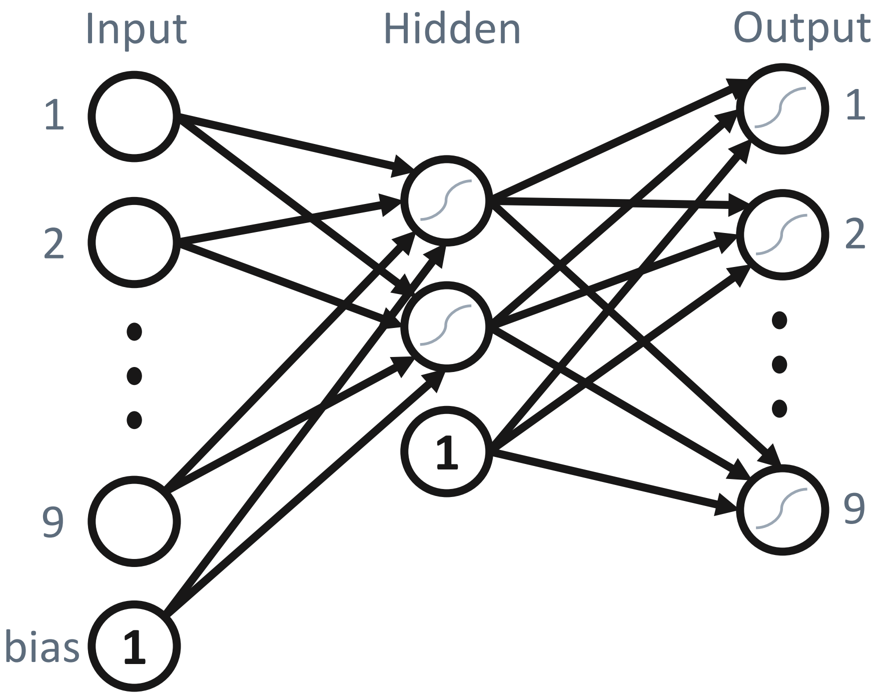


- 255,168 unique games
  - 131,184 are won by the first player
  - 77,904 are won by the second player
  - 46,080 are drawn

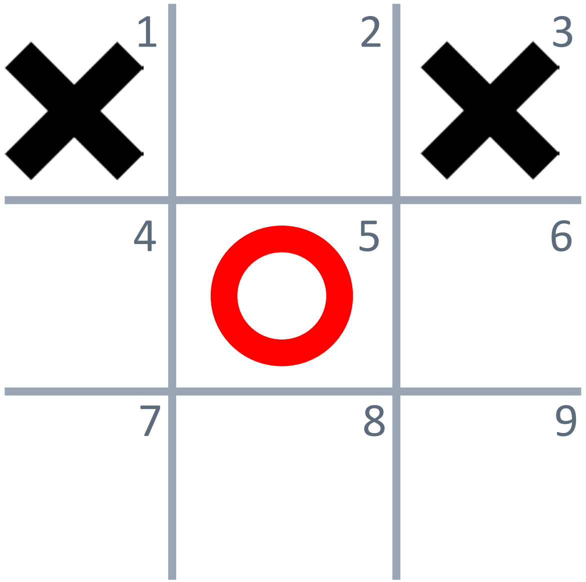
Jesper Juul. “255,168 ways of playing Tic Tac Toe”

Tic Tac Toe





Tic Tac Toe



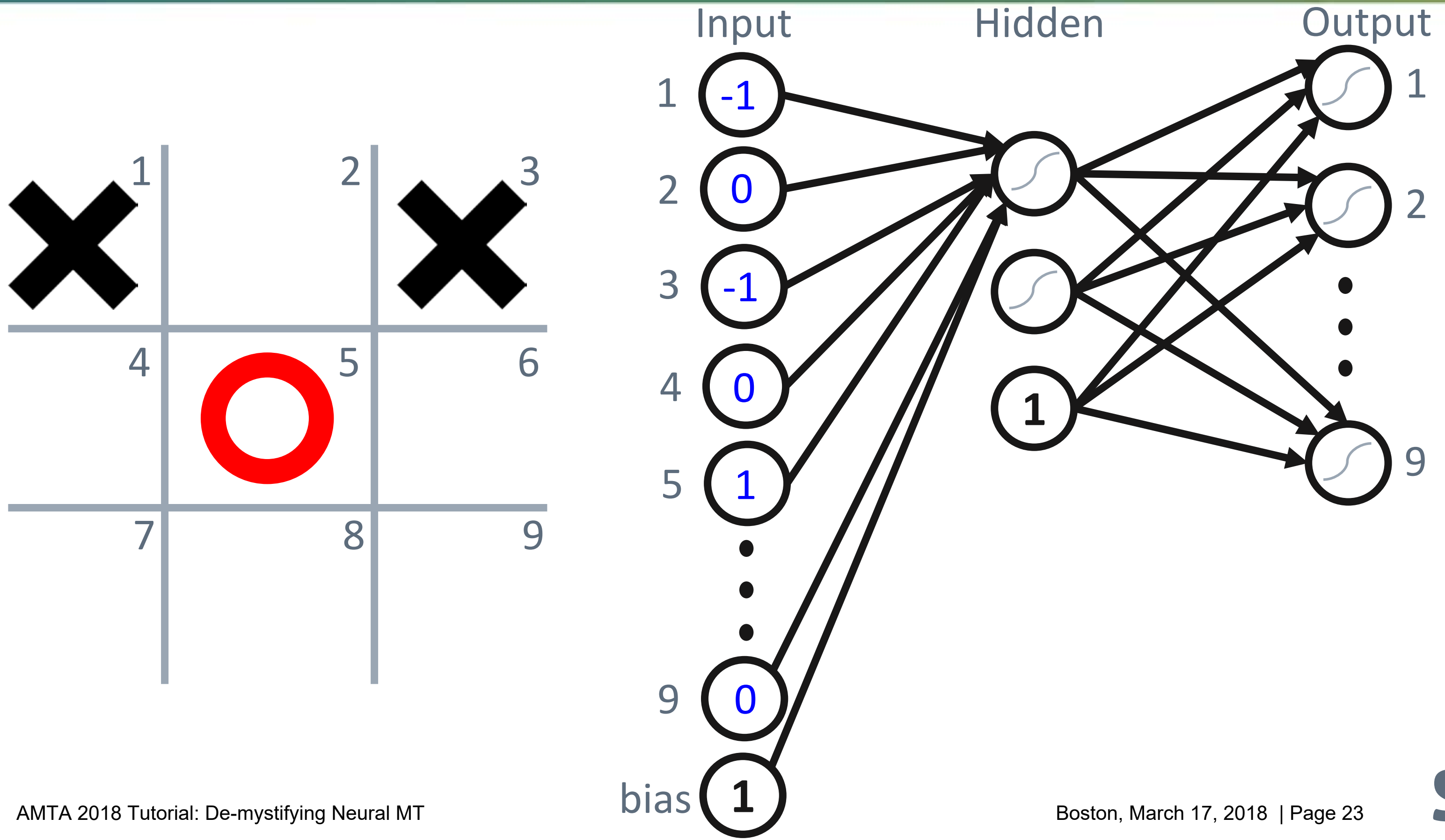
Input representation

- Marked by self: 1
- Marked by opponent: -1
- Empty: 0

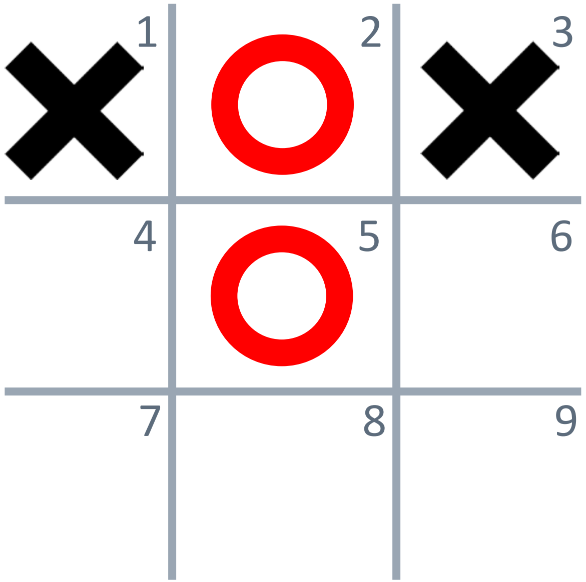
- If computer is O, then:

  [-1, 0, -1, 0, 1, 0, 0, 0, 0]

Tic Tac Toe



Tic Tac Toe



- Input:

  [-1, 0, -1, 0, 1, 0, 0, 0, 0]

- Output:

  [~~0.12~~, **0.8**, ~~0.05~~,
  0.3, ~~0.05~~, 0.37,
  0.41, 0.2, 0.49]

Deep Learning & Deep Neural Networks

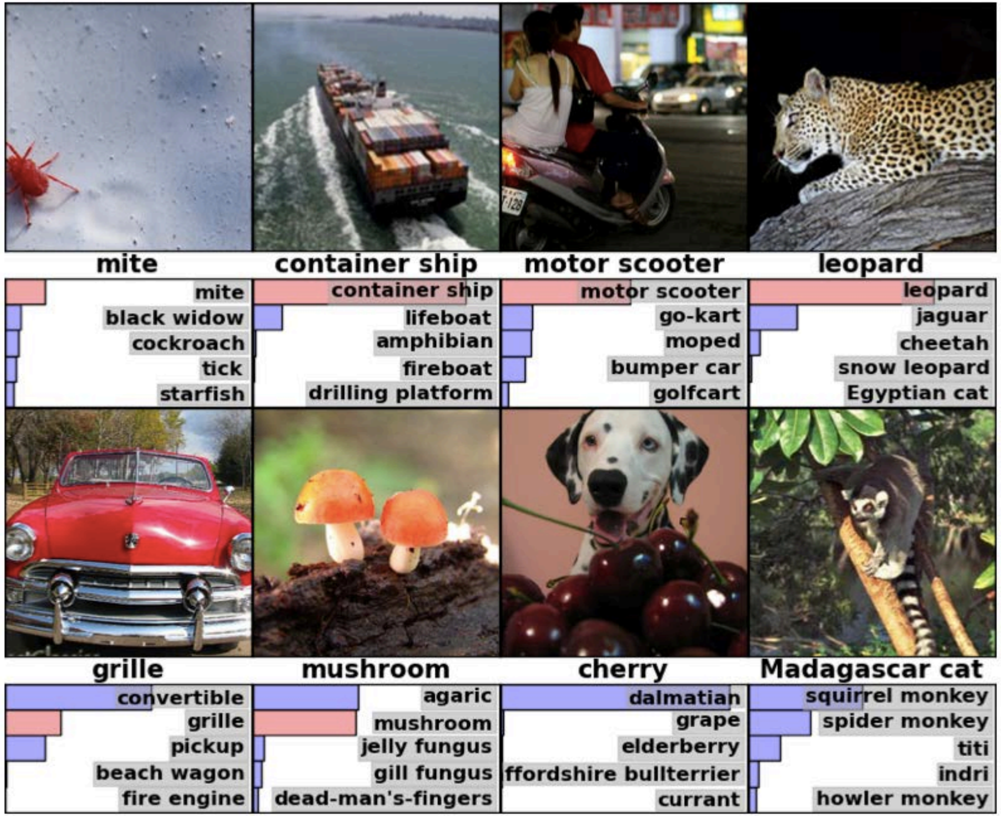
Deep Neural Networks

Multiple Layers

Millions of Parameters

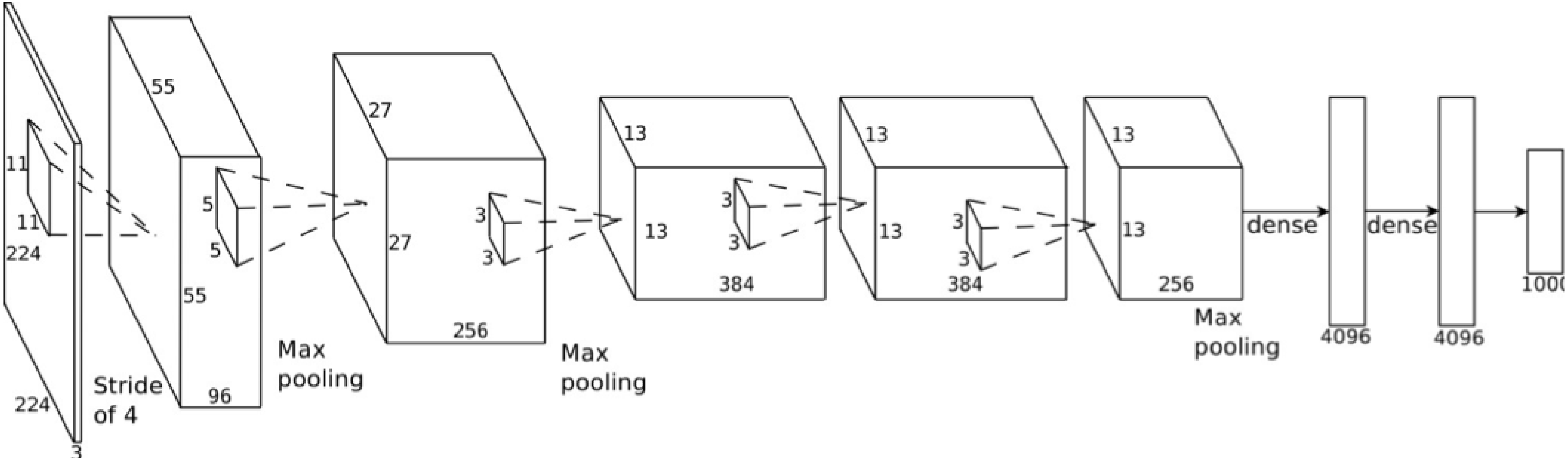
Various Architectures

Deep Neural Network – Image Classification



A Deep Neural Network (convolutional)

cat



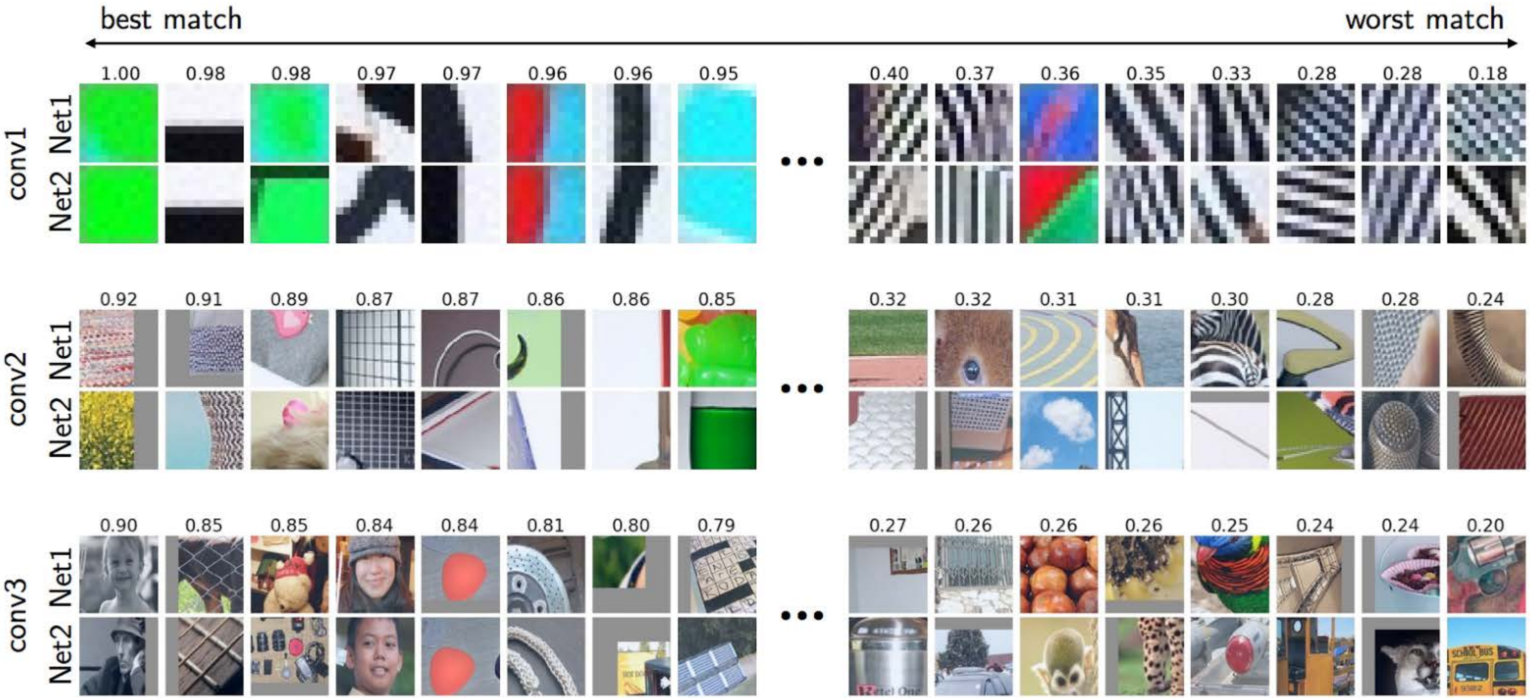
dog

60 million parameters

Why are Deep Networks better?

- Different layers can learn different levels of abstraction
- Mathematically, it can represent more complex functions

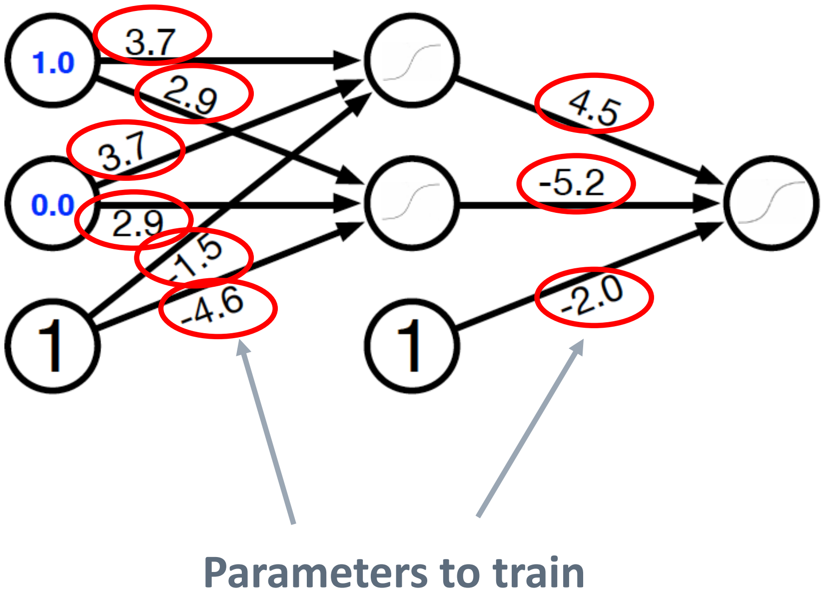
Deep Neural Network – how they learn



Li, Yixuan, et al. "Convergent Learning: Do different neural networks learn the same representations?."

TRAINING

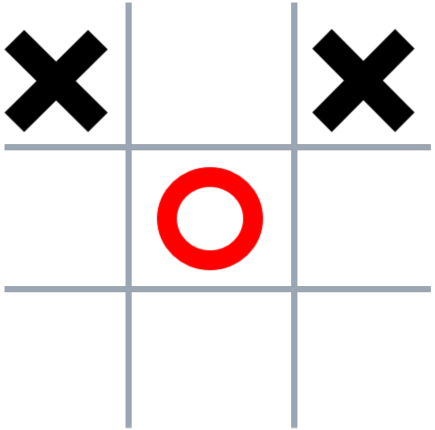
Training: what does a model consist of?

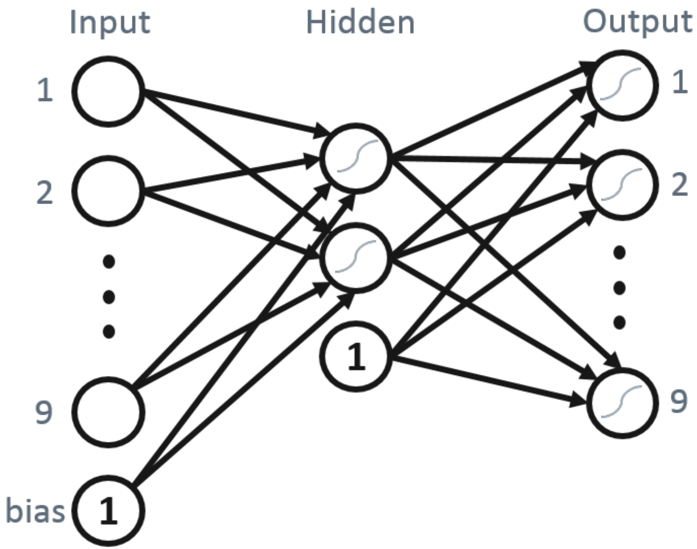


- Each circle with represents an activation function
- Each arrow represents a multiplication
  - input x weight

Training

- What does training actually do?
  - Determine parameter values by minimizing error



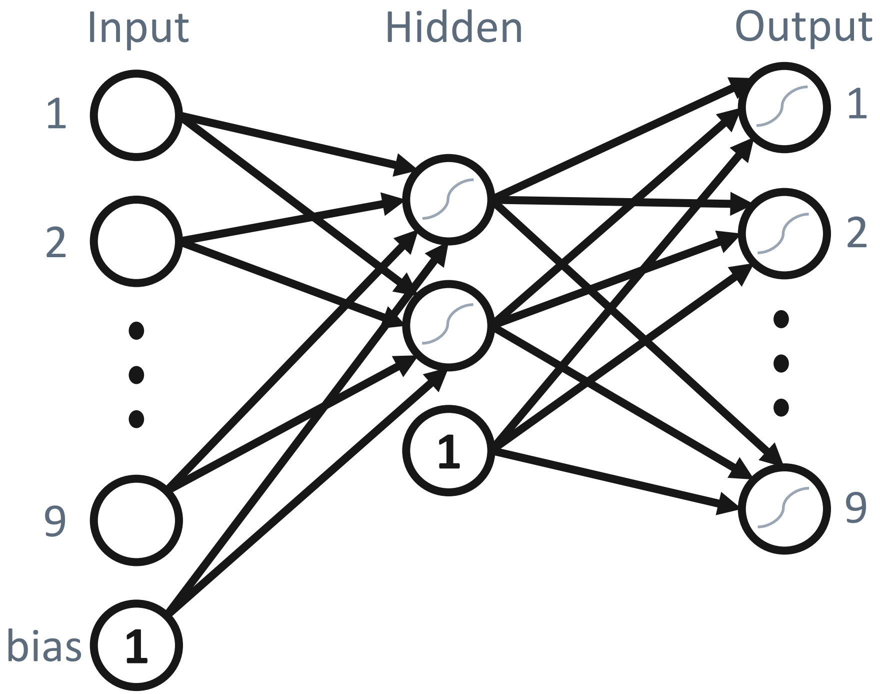


SDL

Training: parameters

- 9 input nodes
- 1 hidden layer: 2 nodes
- 9 output nodes
- Number of parameters

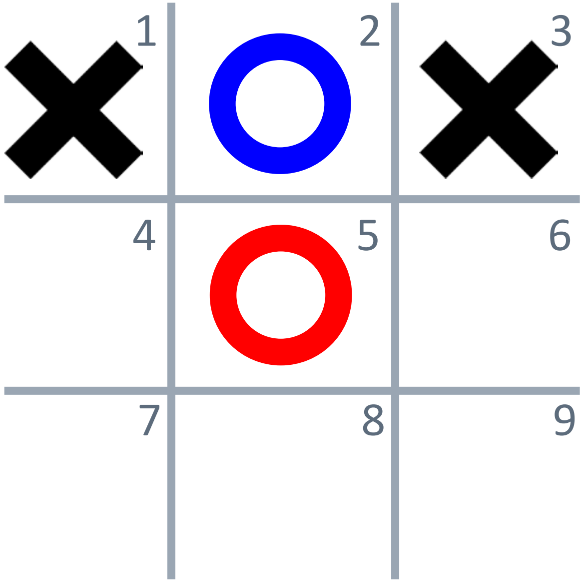
$$= (9 + 1) * 2 + (2 + 1) * 9$$
$$= 47$$



Training

- Steps:
  1. Compute current model output (forward pass) for each training example
  2. Compute cost
  3. Update parameters (backpropagation)

Training: 1. Forward pass



An example of training data

Input
[-1, 0, -1, 0, 1, 0, 0, 0, 0]

Expected output	Model output
[0, **1**, 0, 0, 0, 0, 0, 0, 0]	[.2, .13, .56, .8, .3, .49, .52, .23, .01]

Training: 2. Compute cost

- Cost = error between expected and model output
- Example of a cost function:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

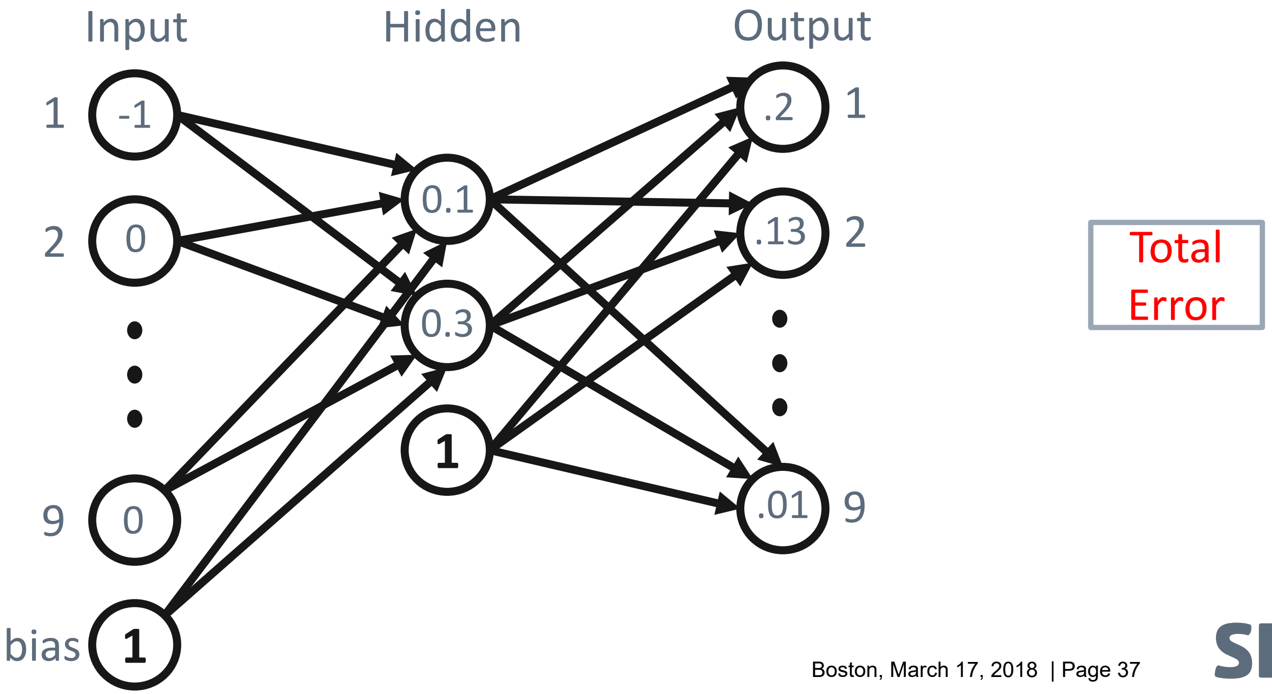
Cost = $\frac{1}{9}((0-.2)^2 + (1-.13)^2 + \cdots)$
= 0.2671

Input
[-1, 0, -1, 0, 1, 0, 0, 0, 0]

Expected output	Model output
[0, **1**, 0, 0, 0, 0, 0, 0, 0]	[.2, .13, .56, .8, .3, .49, .52, .23, .01]

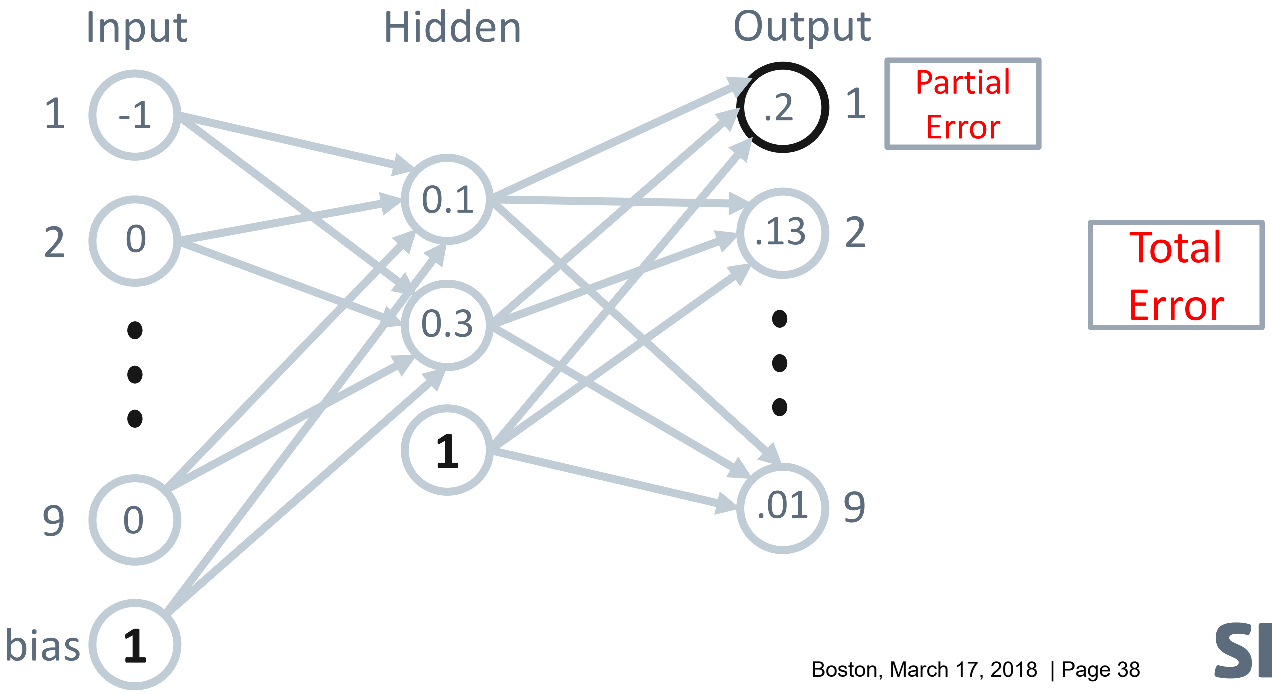
Training: 3. Backpropagation

- Update weights



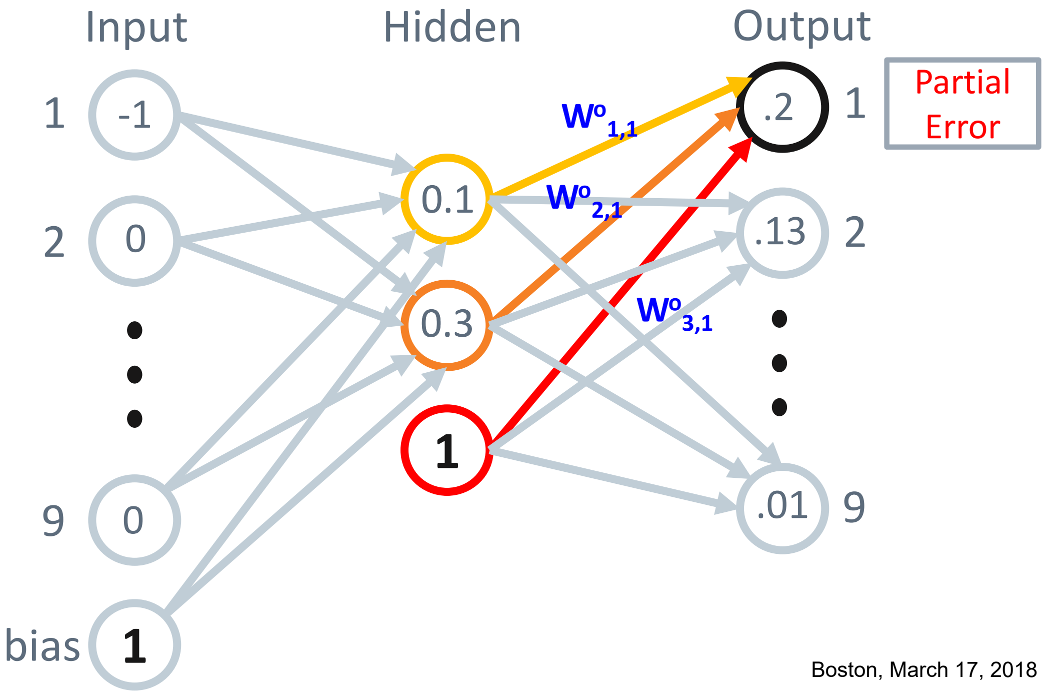
Training: 3. Backpropagation

- Update weights: proportional to activation



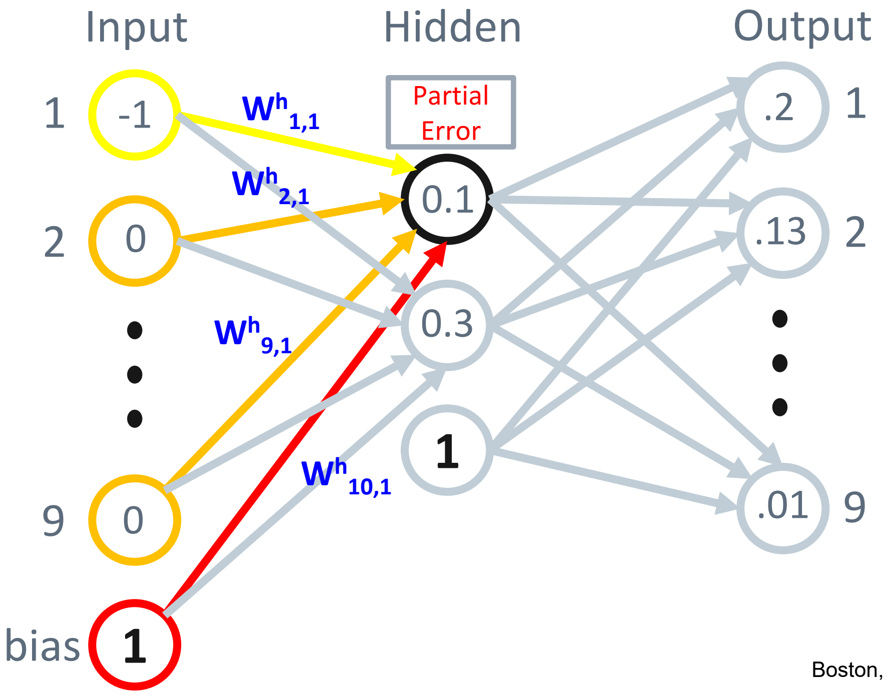
Training: 3. Backpropagation

- Update weights: proportional to activation



Training: 3. Backpropagation

- Update weights: proportional to activation



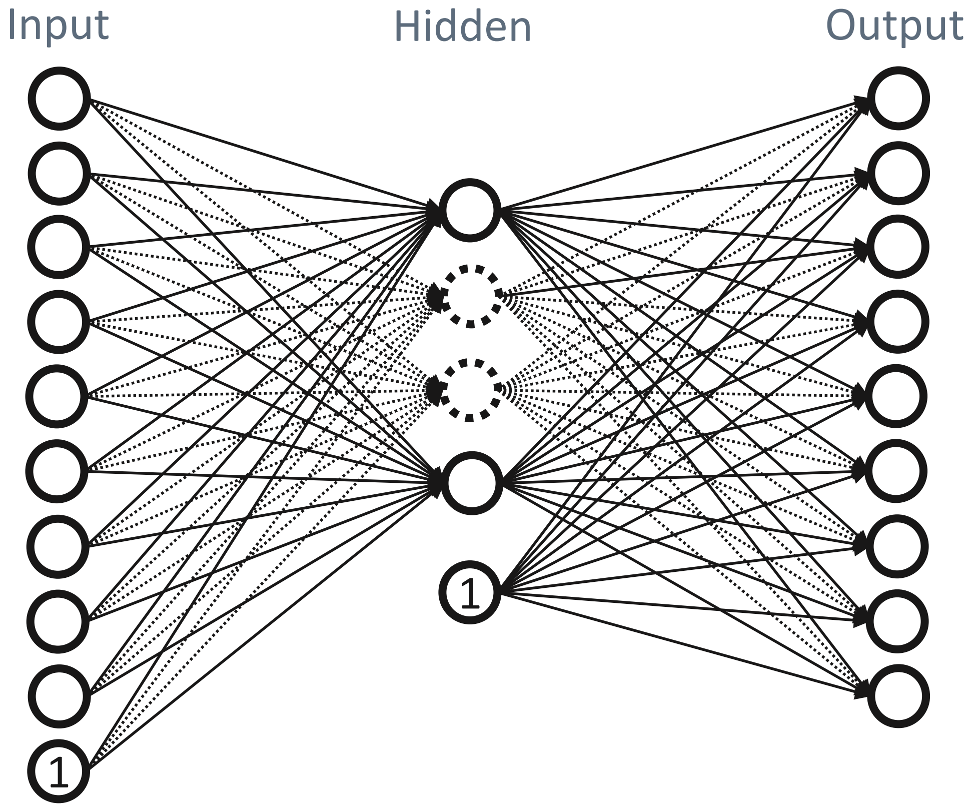
Visualize Neural Network training

- http://www.emergentmind.com/neural-network

Difficulties

- If you just take some data and run backprop, you won't get a good network
  - Especially a deep network
- Some of the problems are:
  - Overfitting
  - Exploding/vanishing gradient

Training tricks: drop-out



Training tricks: synthetic data

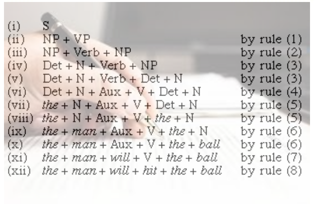
- Increase the amount of data
- Add noise

Oravec, Milos, et al. "Efficiency of recognition methods for single sample per person based face recognition." *Reviews, Refinements and New Ideas in Face Recognition*. InTech, 2011.

MACHINE TRANSLATION

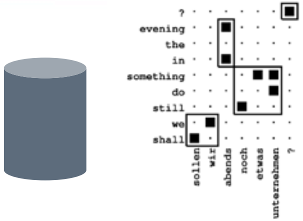
Machine Translation

1949



RBMT

1970s



SMT Phrase-Based

2001

Adaptive MT

Neural MT

2017

1966

ALPAC

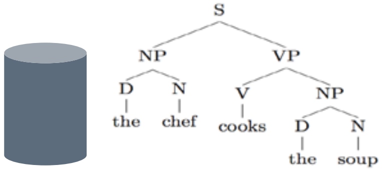
Report

1993

$$\tilde{e} = arg \max_{e \in e^*} p(e|f)$$

SMT

2008



SMT Syntax-Based

Statistical Machine Translation

Translation Model

Language Model

Syntax Model

Capitalization

Distortion

Morphology

Alignment

Transliteration

Part of Speech

Smoothing

Reordering

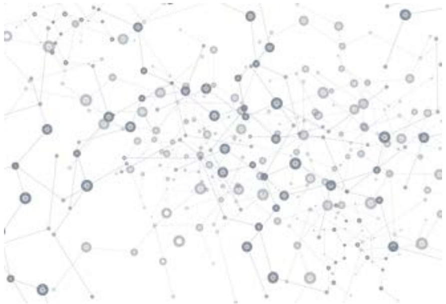
Word Deletion

Lexicalized Reordering

Preordering

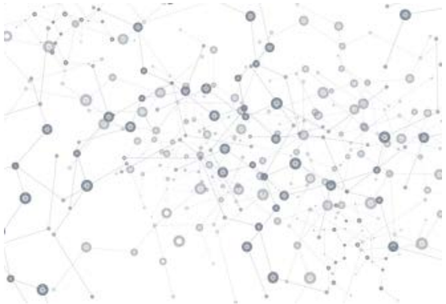
Neural Machine Translation

Input
Text



ENCODER

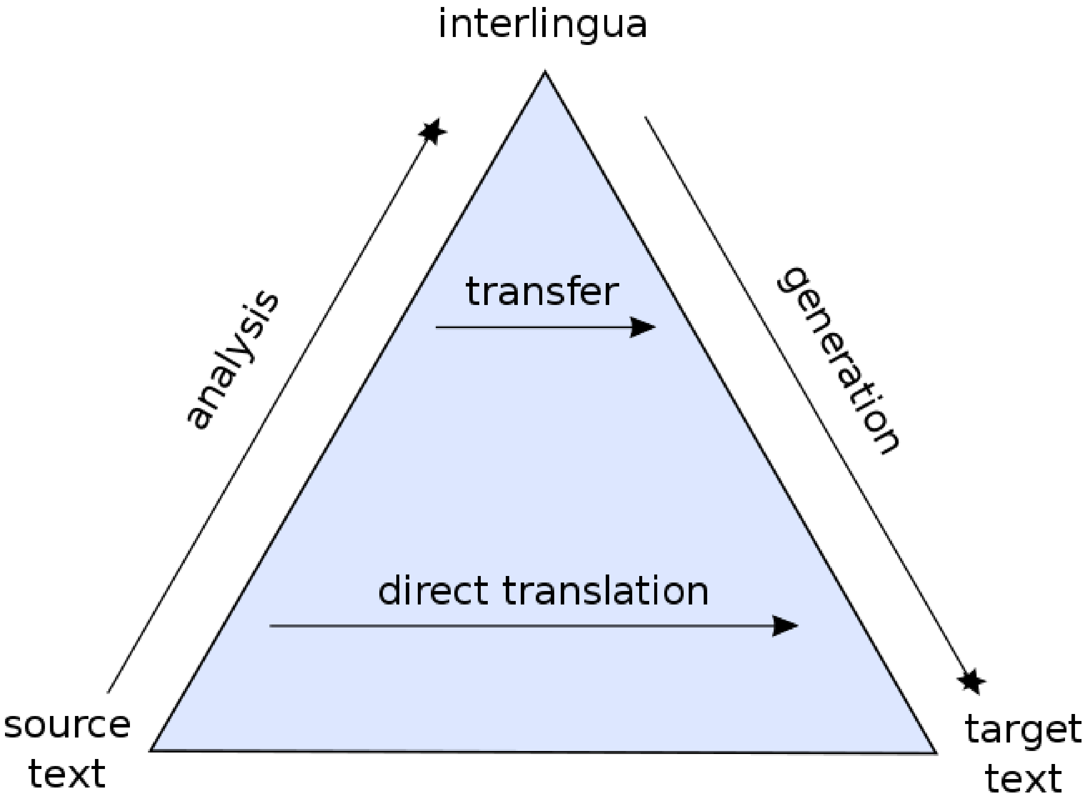
-0.2
-0.1
0.1
0.4
-0.3
1.1
4.3
-0.2
0.5
0.9
1.3
3.4
-5.3
-6.2
4.8
9.3
3.4
…
2.6
4.9
0.1
2.6
8.3
-7.3
5.1
1.5
0.6
9.3
-6.2
2.9
1.4
-1.3



DECODER

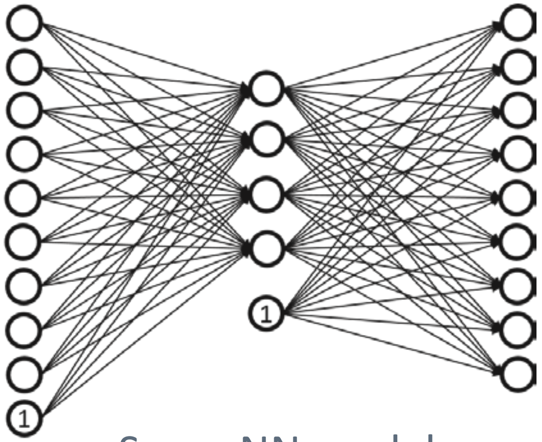
Output
Text

The Machine Translation pyramid



NMT: Word Representations

"My name is Dragos"



Some NN model

"Je m'appelle Dragos"

NMT: Word Representations – one hot

Vocab	1	2	3	4	5	6	7	8	9	10
a	1	0	0	0	0	0	0	0	0	0
burger	0	1	0	0	0	0	0	0	0	0
dragos	0	0	1	0	0	0	0	0	0	0
for	0	0	0	1	0	0	0	0	0	0
had	0	0	0	0	1	0	0	0	0	0
i	0	0	0	0	0	1	0	0	0	0
is	0	0	0	0	0	0	1	0	0	0
lunch	0	0	0	0	0	0	0	1	0	0
my	0	0	0	0	0	0	0	0	1	0
name	0	0	0	0	0	0	0	0	0	1

NMT: Word Representations – one hot

- "my name is dragos"

  Index: [9, 10, 7, 3]

  One-hot:

  – "my" (9): [0, 0, 0, 0, 0, 0, 0, 0, 1, 0]

  – "name" (10): [0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

  – "is" (7): [0, 0, 0, 0, 0, 0, 1, 0, 0, 0]

  – "dragos" (3): [0, 0, 1, 0, 0, 0, 0, 0, 0, 0]

NMT: Word Representations – one hot

- Problem with this method:
  - Large number of vocab => curse of dimensionality
  - Hard to capture the relationships between words

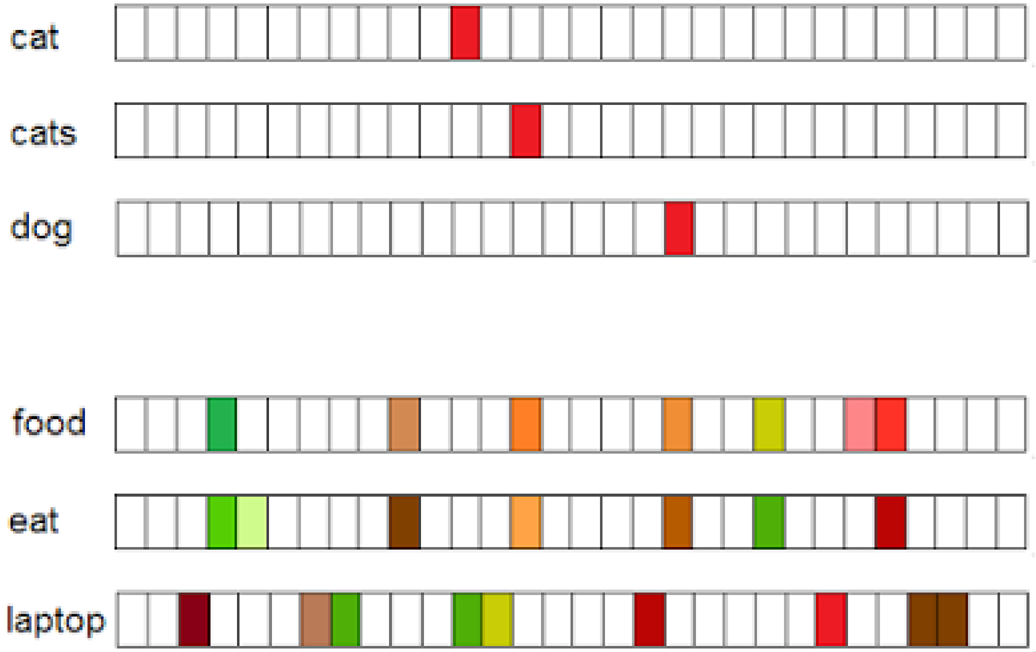
Word representations

Sparse

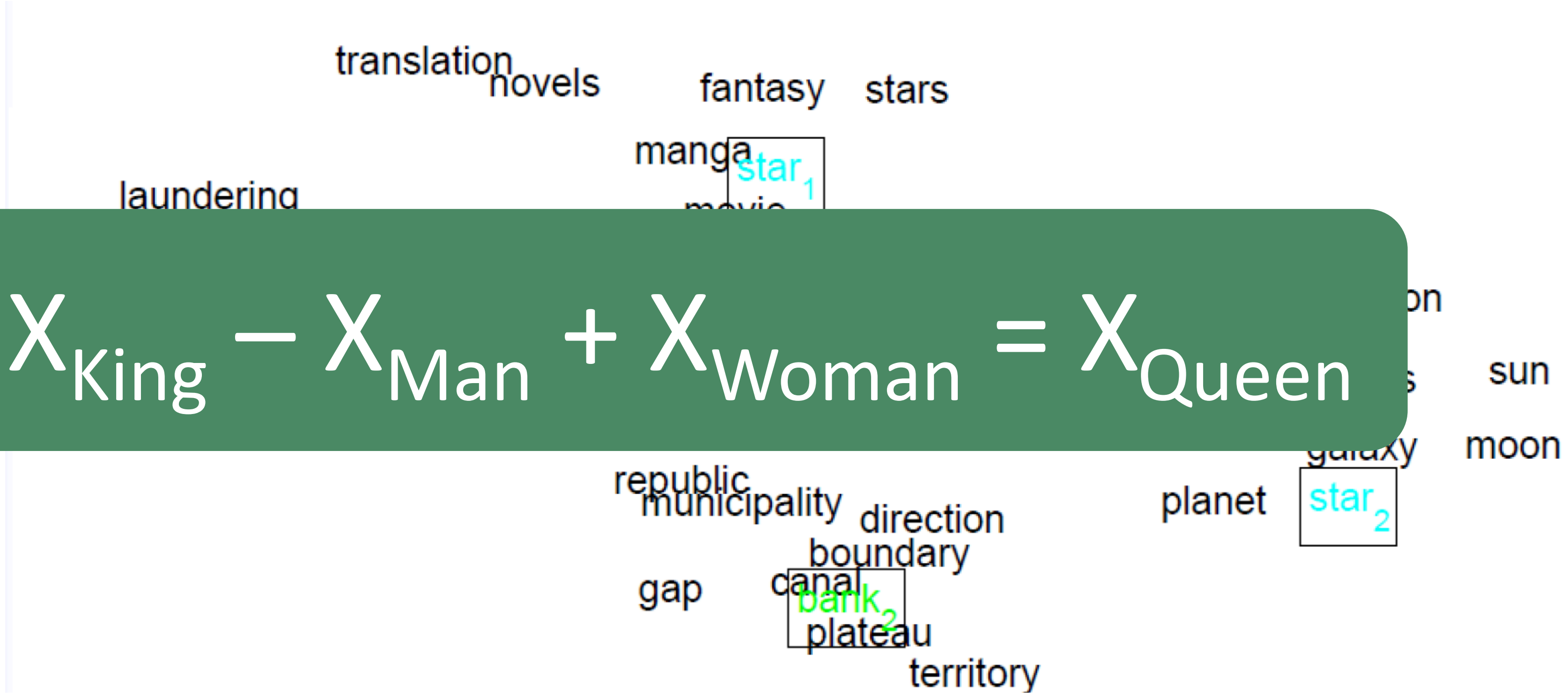
All words are equally different

Dense

Similar words have similar vectors



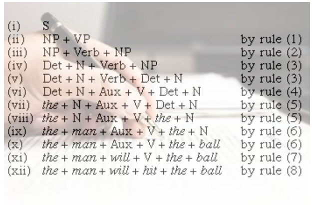
NMT: Word Representations and Word Embedding



$$X_{King} - X_{Man} + X_{Woman} = X_{Queen}$$

BREAK

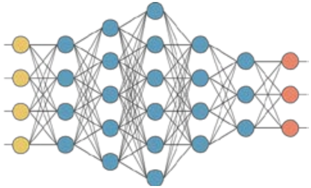
Rule-based vs. Statistical vs. Neural



Rule-Based

$$\tilde{e} = arg\max_{e \in e^*} p(e|f)$$

Statistical



Neural

Statistical Machine Translation

- Translation Model
- Language Model
- Distortion
- Alignment
- Part of Speech
- Reordering
- Lexicalized Reordering
- Syntax Model
- Capitalization
- Morphology
- Transliteration
- Smoothing
- Word Deletion
- Preordering

Neural Machine Translation

Input Text

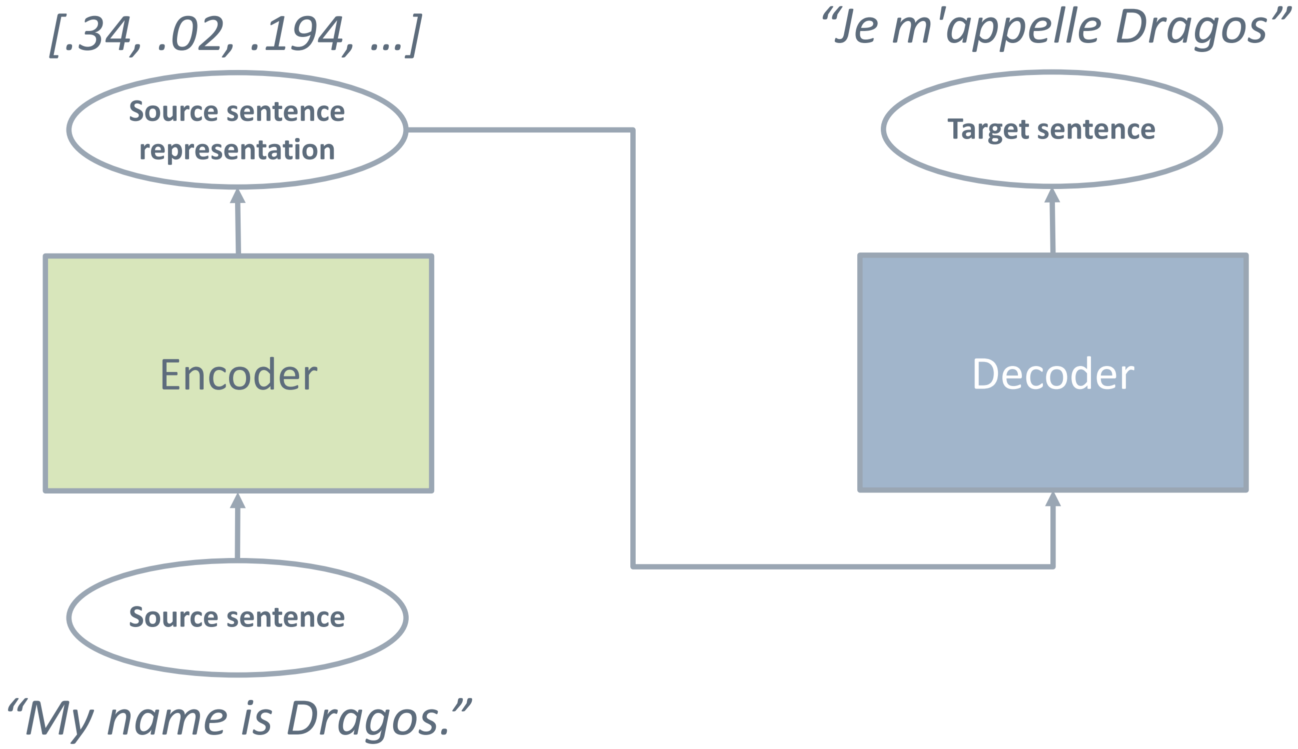
ENCODER

-0.2
-0.1
0.1
0.4
-0.3
1.1
4.3
-0.2
0.5
0.9
1.3
3.4
-5.3
-6.2
4.8
9.3
3.4
…
2.6
4.9
0.1
2.6
8.3
-7.3
5.1
1.5
0.6
9.3
-6.2
2.9
1.4
-1.3

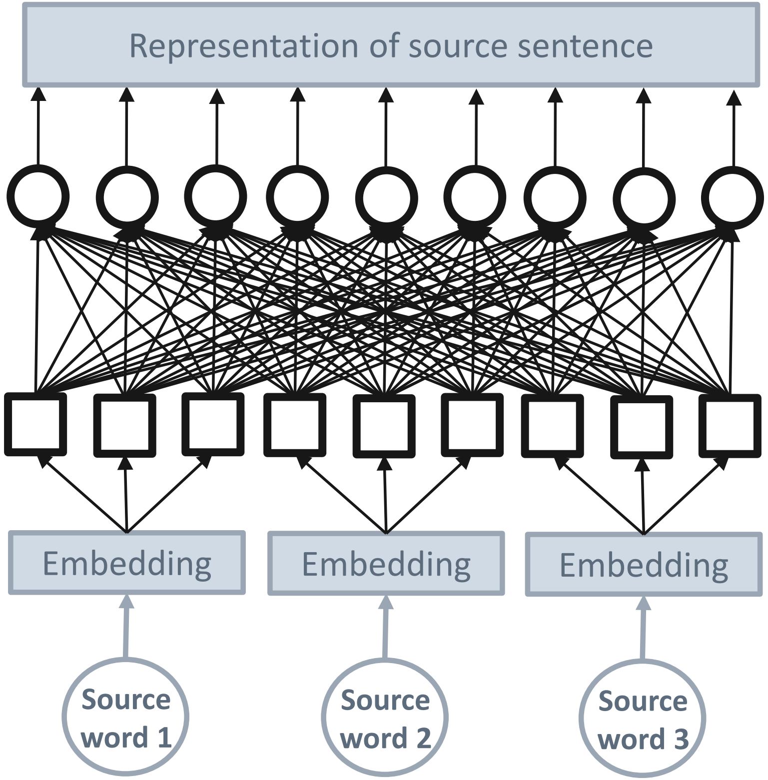
DECODER

Output Text

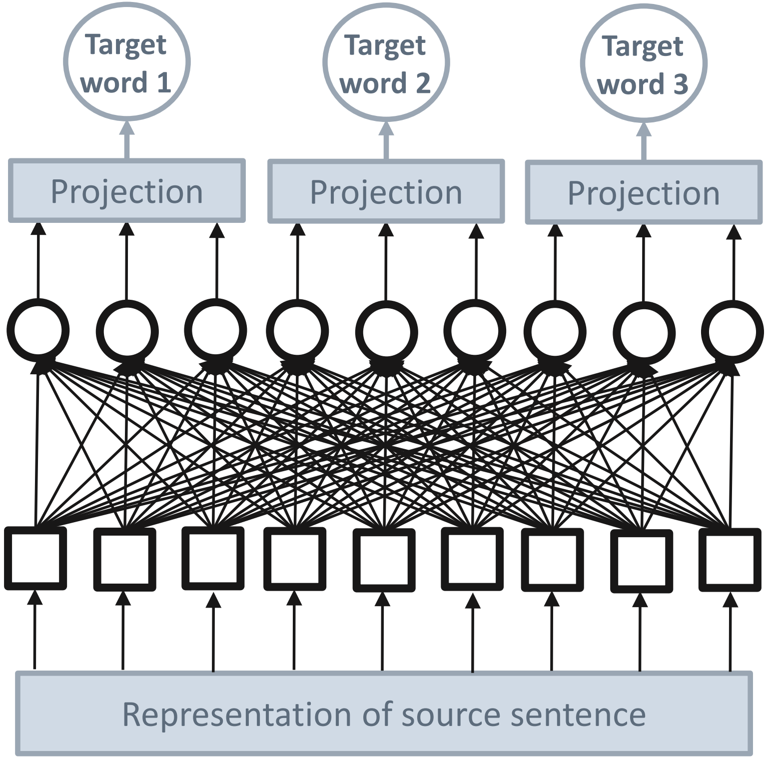
Encoder Decoder



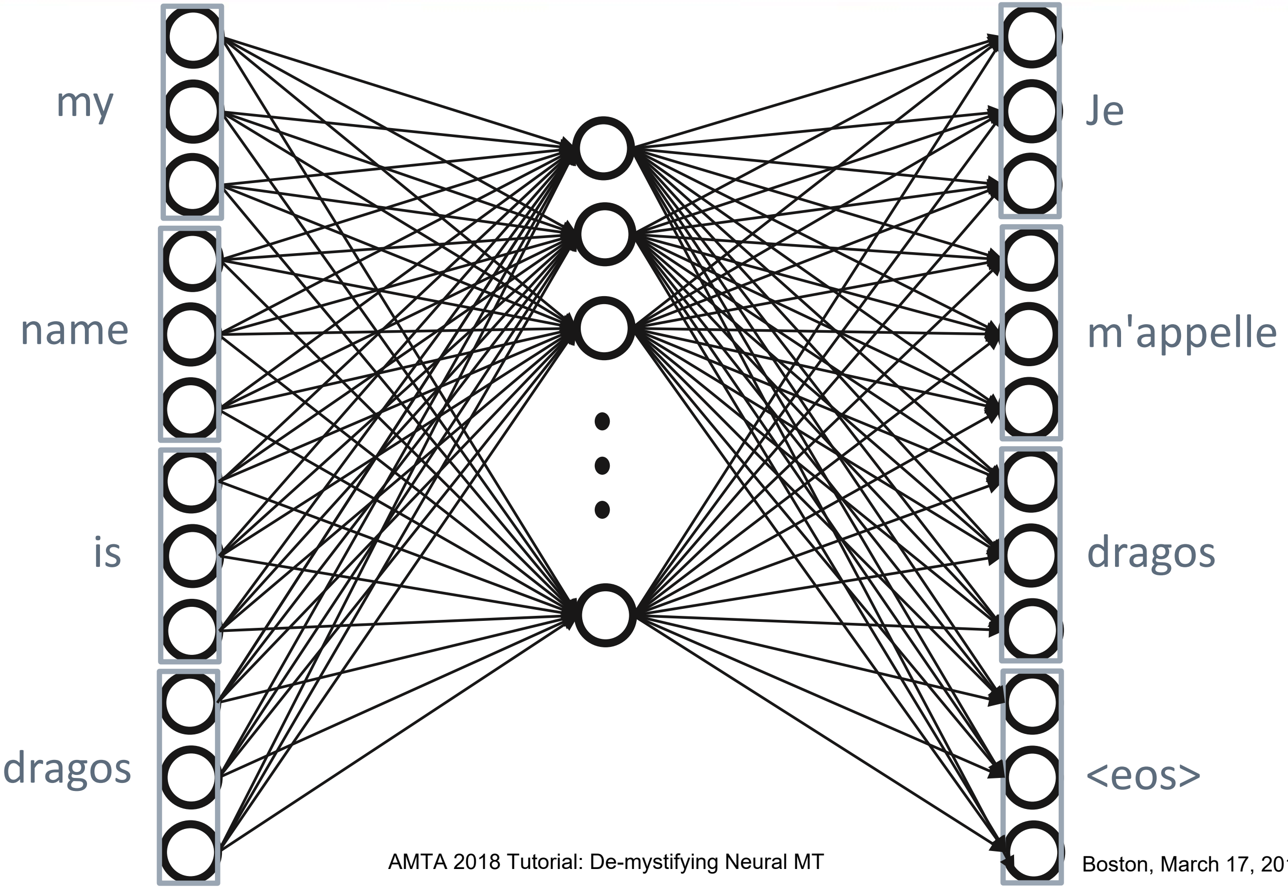
Sequence-to-sequence learning: Encoder



Sequence-to-sequence learning: Decoder



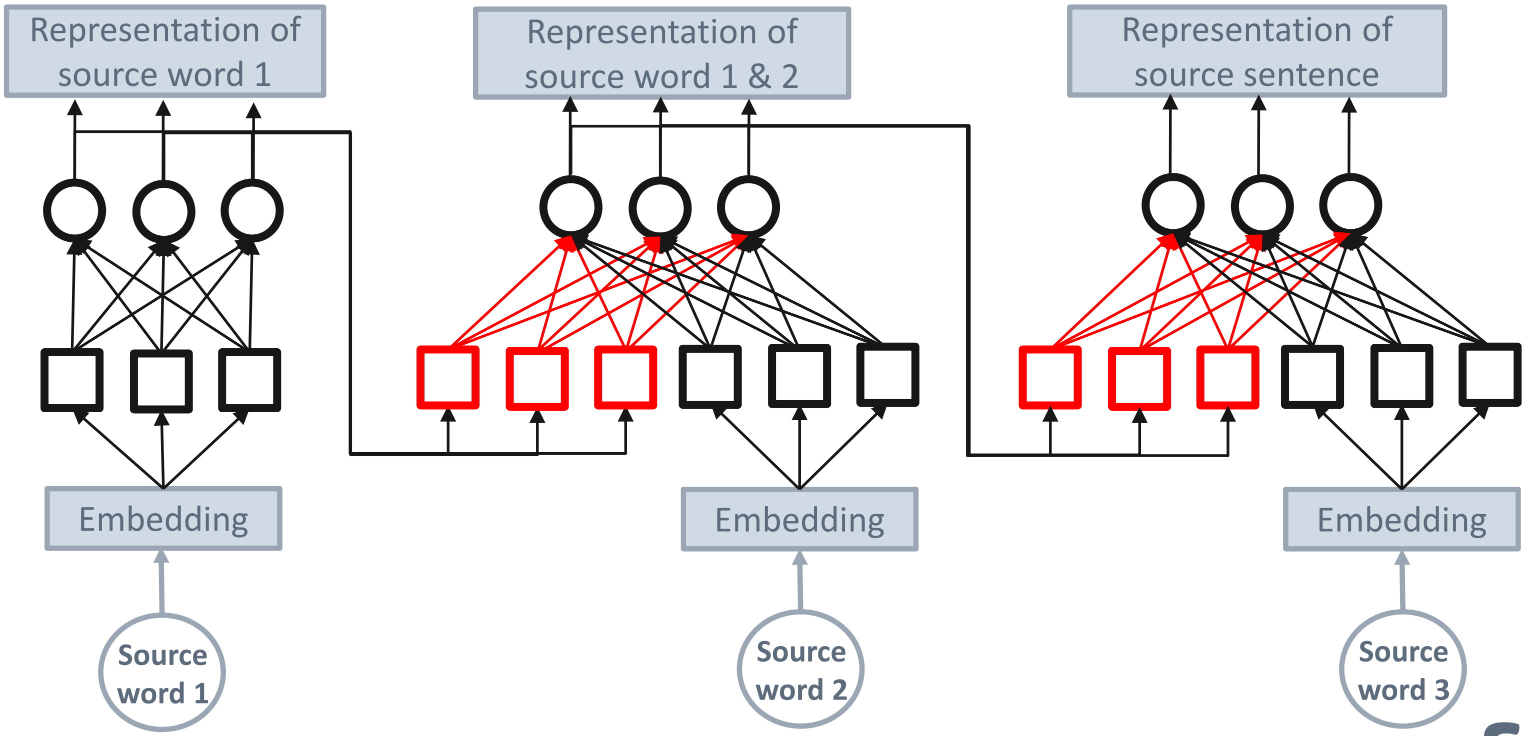
Let’s use a simple NN for machine translation



Sequence-to-sequence learning

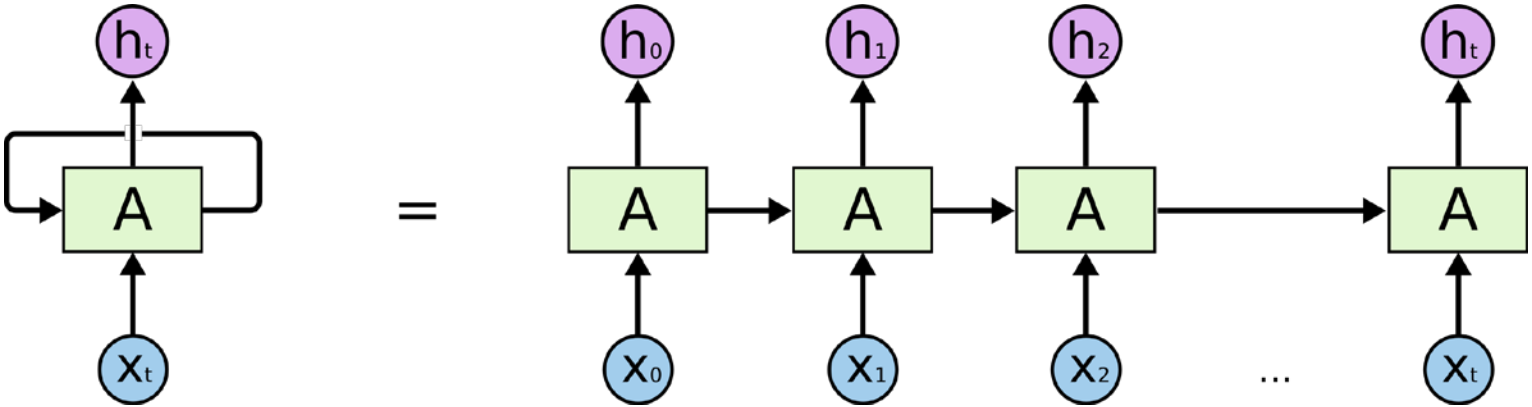
- Example sentences
  - “My name is Dragos.”
  - “Machine translation, sometimes referred to by the abbreviation MT (not to be confused with computer-aided translation, machine-aided human translation (MAHT) or interactive translation) is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one language to another.” [Wikipedia]

Vanilla Recurrent Network



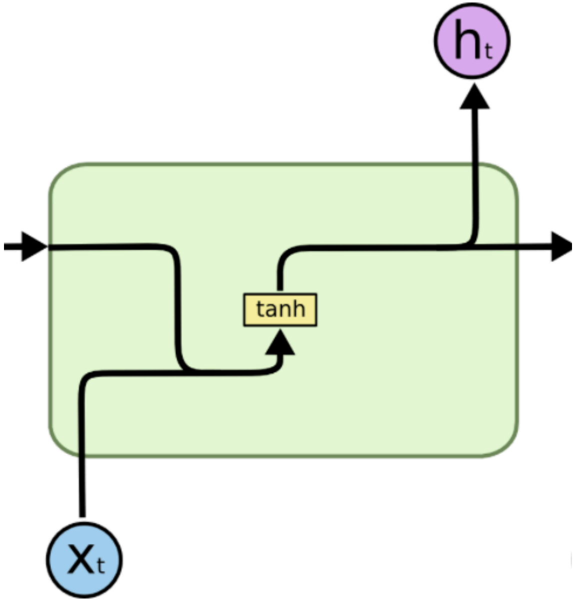
SDL

Recurrent Neural Network

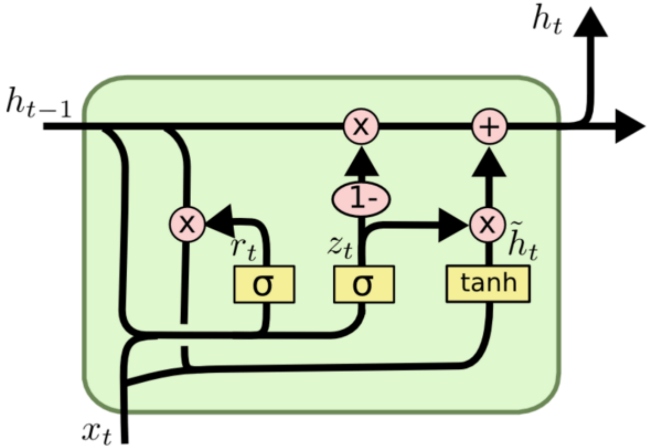


http://colah.github.io/posts/2015-08-Understanding-LSTMs/

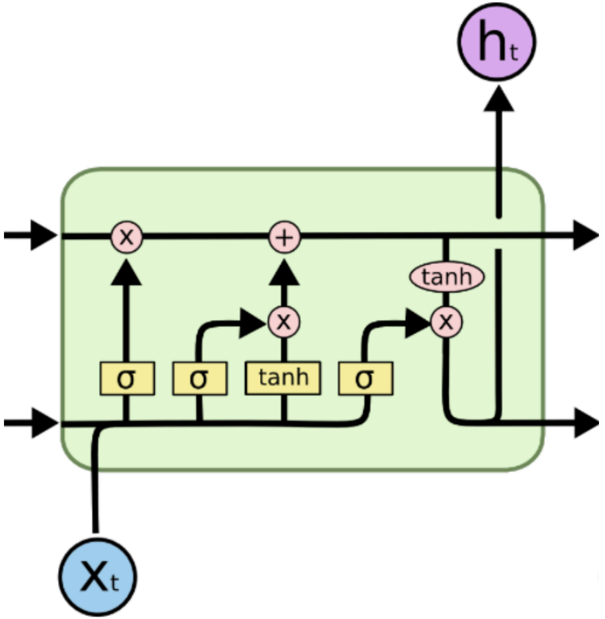
Different RNN units



Vanilla



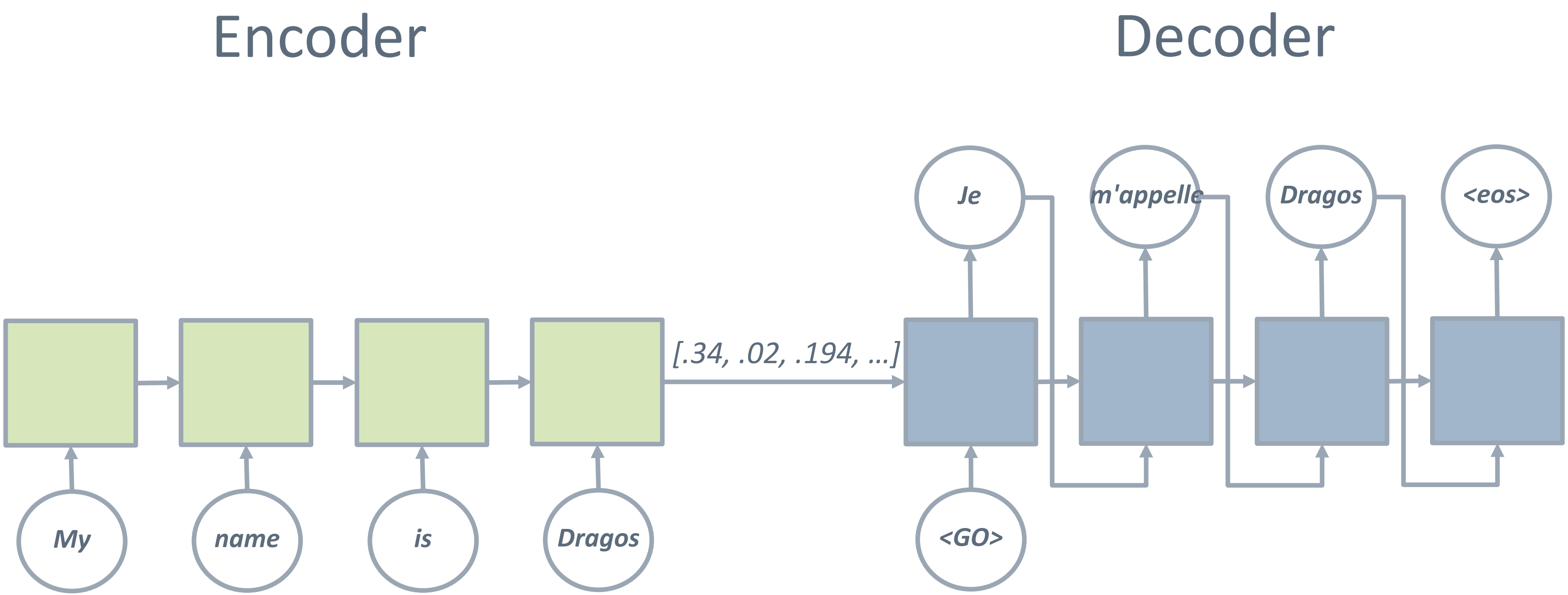
GRU



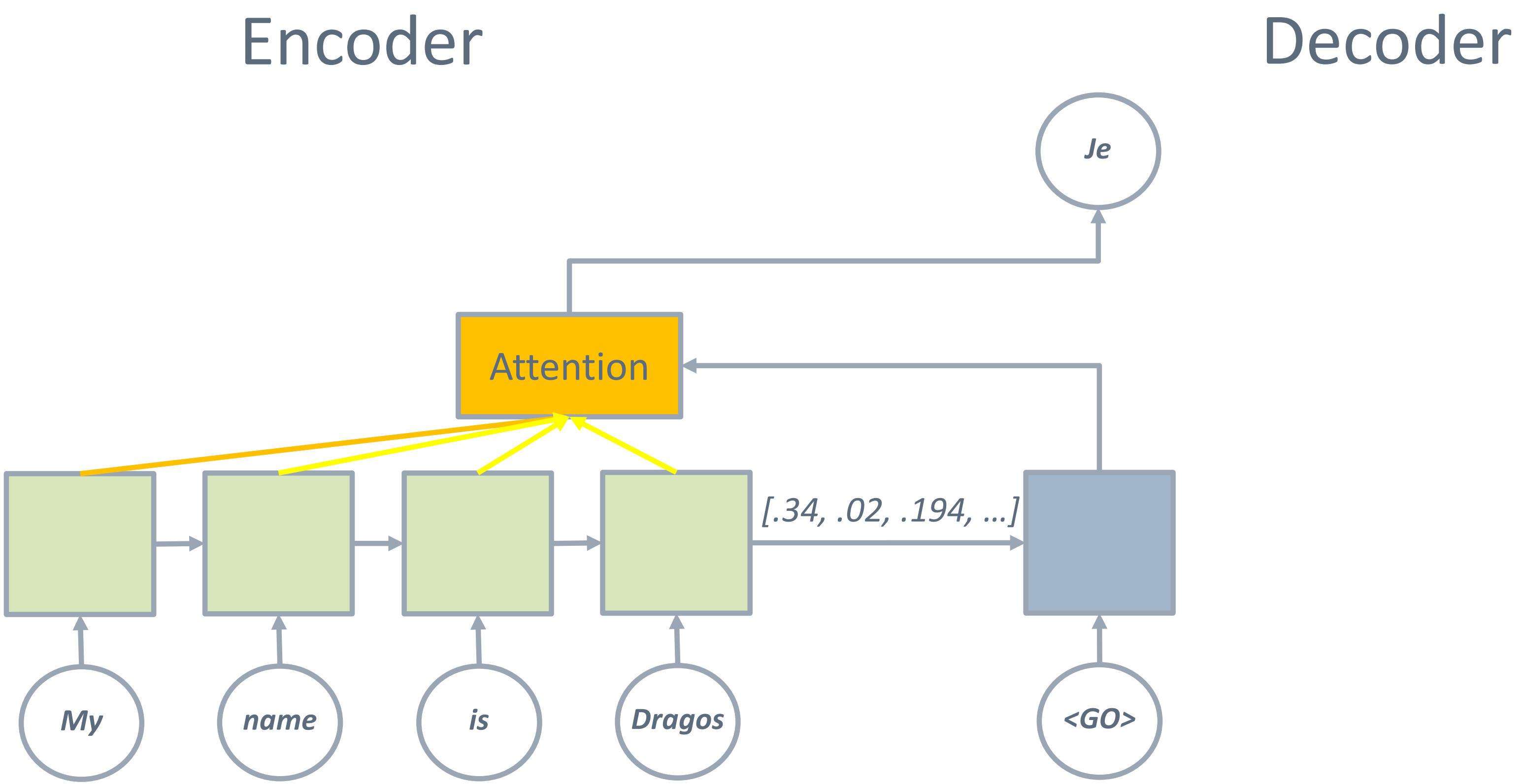
LSTM

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

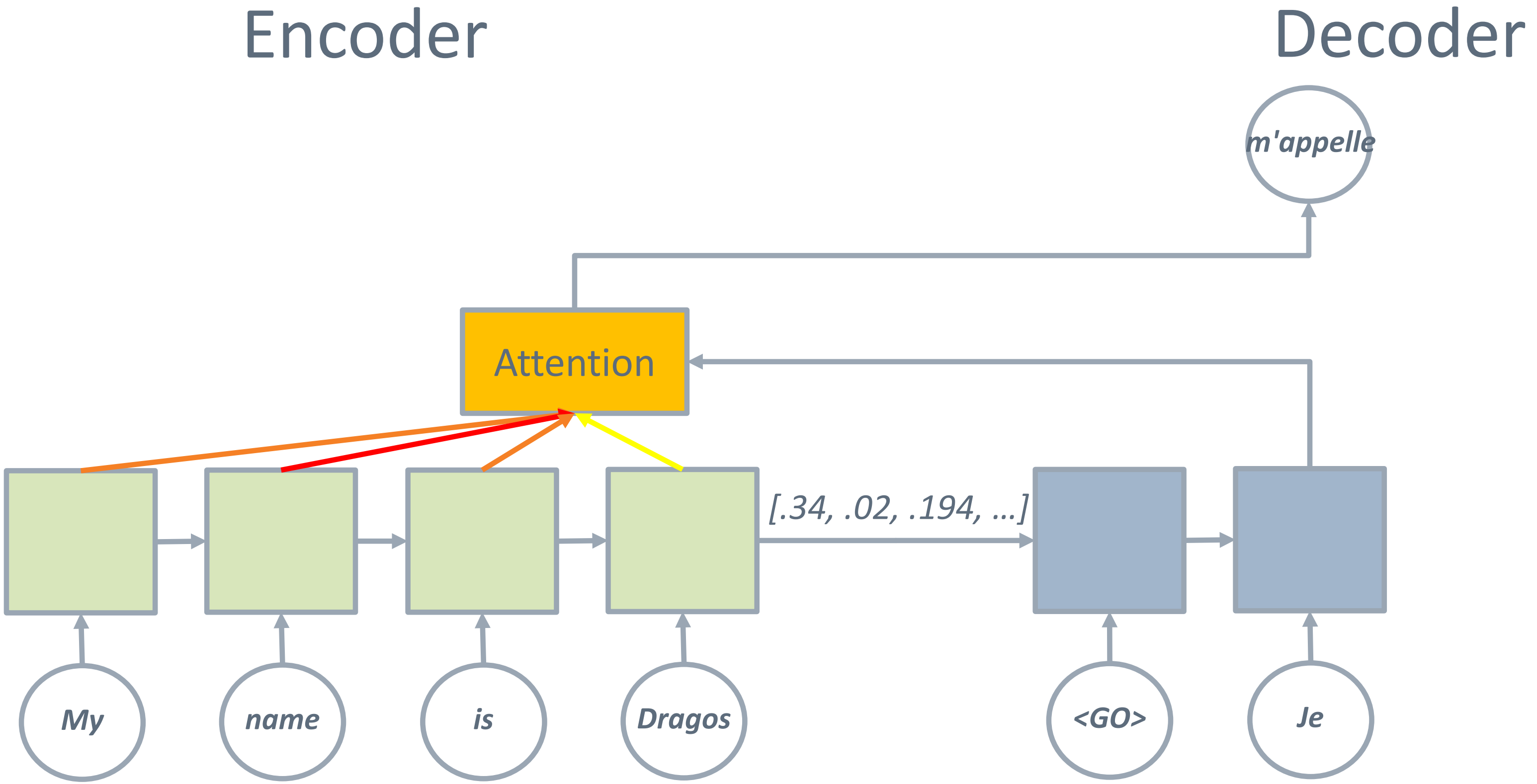
Encoder Decoder unrolled



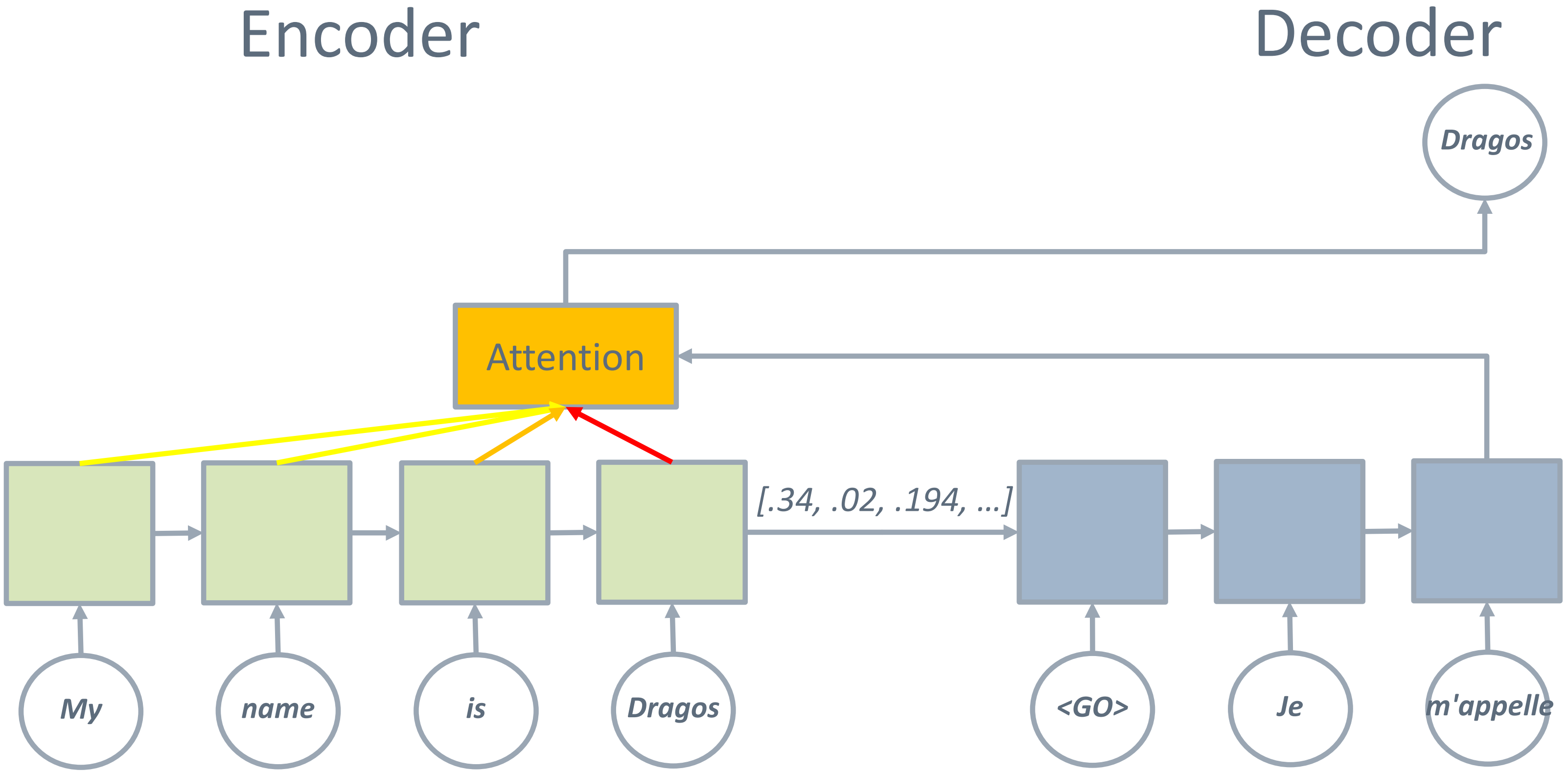
Encoder Decoder with Attention



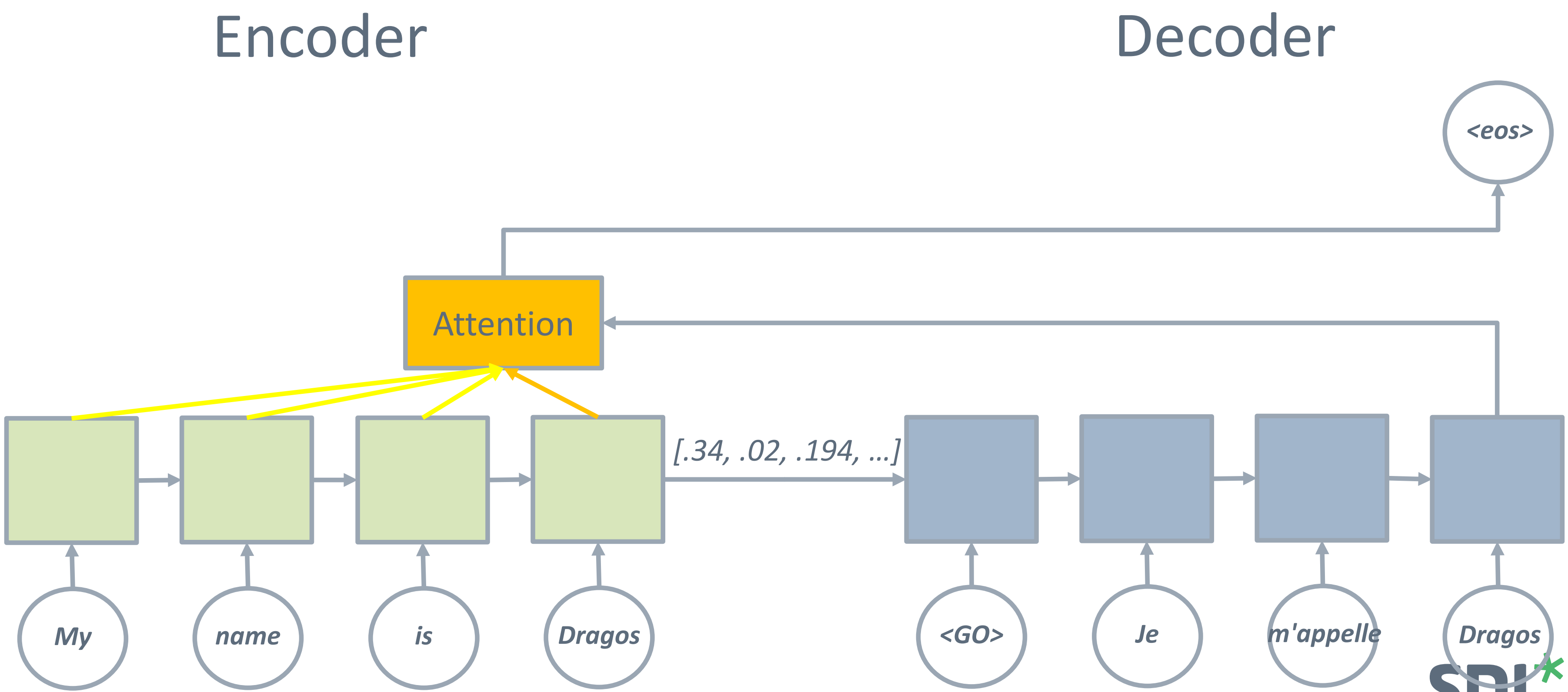
Encoder Decoder with Attention



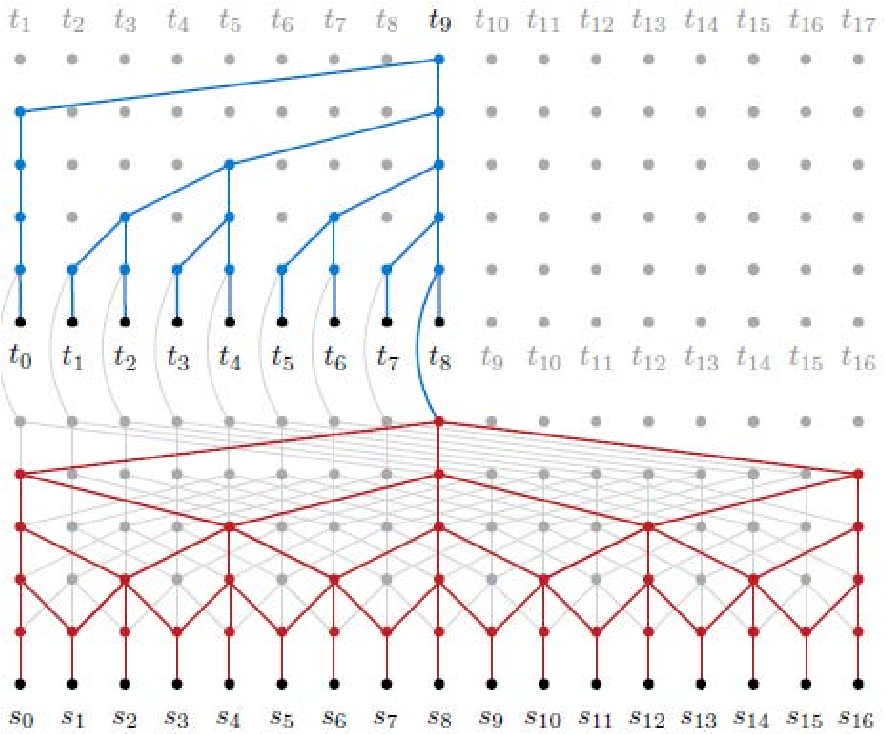
Encoder Decoder with Attention



Encoder Decoder with Attention



Convolutional Model for Machine Translation



Kalchbrenner, Nal, et al. "Neural machine translation in linear time." *arXiv preprint arXiv:1610.10099* (2016).

Transformer Model

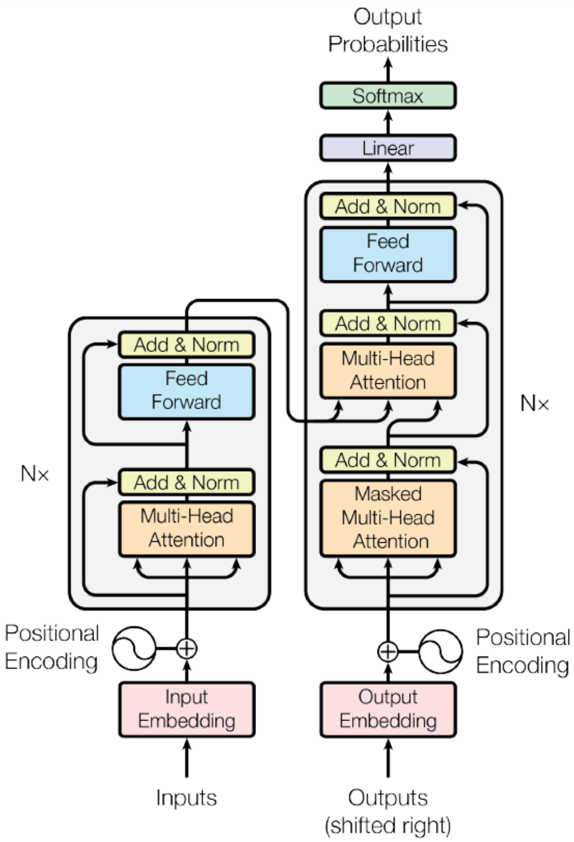


Figure 1: The Transformer - model architecture.

- No recurrence
- No convolution
- More parallelizable
- Use self-attention

Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.

Winograd schema sentences

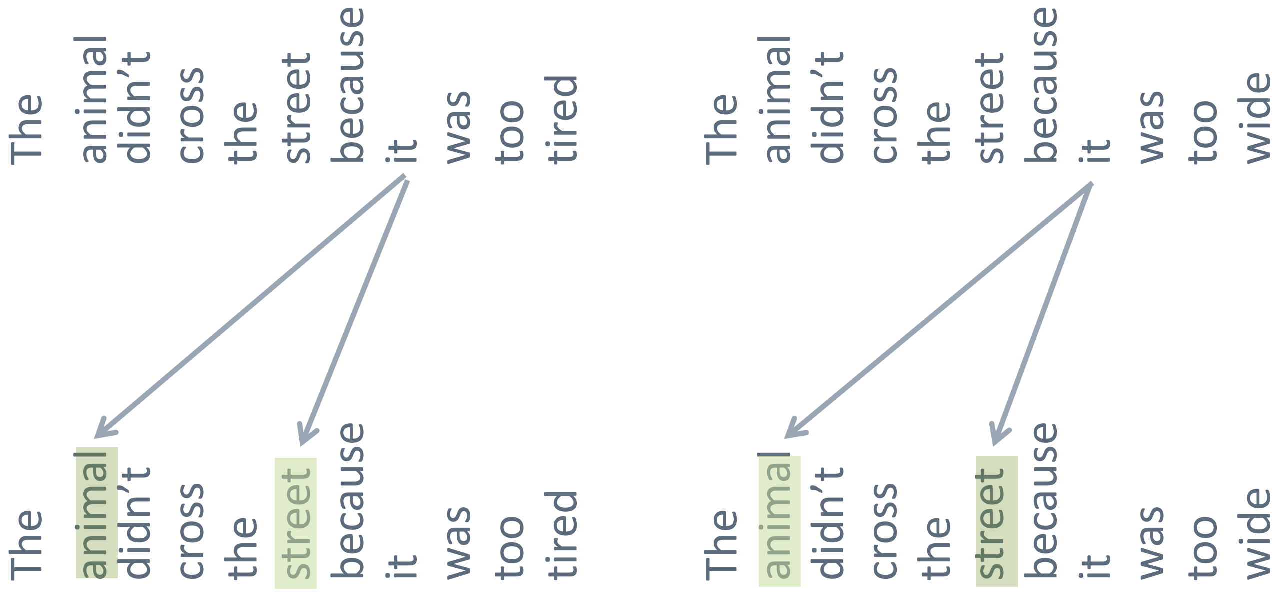
He didn’t put the trophy in the suitcase because it was too **small**.
He didn’t put the trophy in the suitcase because it was too **big**.

The cow ate the hay because it was **delicious**.
The cow ate the hay because it was **hungry**.

The councilmen refused the demonstrators a permit because they **advocated** violence.
The councilmen refused the demonstrators a permit because they **feared** violence.

The animal didn’t cross the street because it was too **tired**.
The animal didn’t cross the street because it was too **wide**.

Self-Attention for coreference resolution



Limitations of NMT

- Unseen words
  - Sentence: *“I had a hamburger for lunch”*
  - The model: *“I had a UNK for lunch”*
  - *Sentence: "I don't like rollercoasters"*
  - The model: "*I don't like UNK*"
- Solutions
  - Subword
  - Dictionary ← Not so simple

Limitations of NMT

- Subword
  - *"ham"+"burger" => "hamburger"*
  - *"roll" + "er" + "coast" + ers" => "rollercoasters"*
  - *"d" + "r" + "a" + "g" + "o" + "s" => "Dragos"*

Limitations of NMT

- Resource requirements
  - Large amount of data
  - GPU
- User constraints: names, numbers, terminology
- Coverage
  - Dropping translation

Limitations of NMT

- Neurobabble
  - "if you do not have any questions , please do not have any questions"
  - "in the middle of the middle of the river , the river flows into the south of the river"
  - "... confronting the history of the history of the history of the history"

Advantages of NMT

- Example:
  - "因此，要改善机器翻译的结果，人为的介入仍显相当重要。"
  - Literal: "Therefore, to improve machine translation results, human intervention is still very important“
  - SMT: "Therefore, it is necessary to improve machine translation results, human intervention is still the video was important."
  - NMT: "Therefore, human intervention is still significant in order to improve the results of machine translation."

Future Outlook

- Adaptation
- Low-resource languages
- Multi-lingual models
- Multi-modal models (speech, image, etc.)

QUESTIONS & ANSWERS