DEEP LEARNING ON NLP (1)

Lili Jiang, Xuan-Son Vu , Michele Persiani

Organised by HPC2N

08 May, 2019

UMEÅ UNIVERSITY

- 13:00-14:00 Lecture: Introduction of Deep Learning on Natural Language Processing (NLP).

- 14:00-14:30 Exercise: Running data representation in basic (RNN).

14:30 Coffee break

- 15:00-16:00 Exercise: Play with different state-of-the-art deep learning models on a NLP application (e.g., Transformers)

COURSE WEBSITES

- Course github

https://github.com/ddmatumu/TFnDeepLearning

- HPC2N website

https://www.hpc2n.umu.se/events/courses/deep-learning-spring-2019

- Feedback:

www.menti.com (Code: 21 40 80)

NATURAL LANGUAGE PROCESSING

- **NLP** deals with building computational algorithms to automatically analyze and represent human language.
  - Google search engine
  - machine translation
  - dialogue generation.

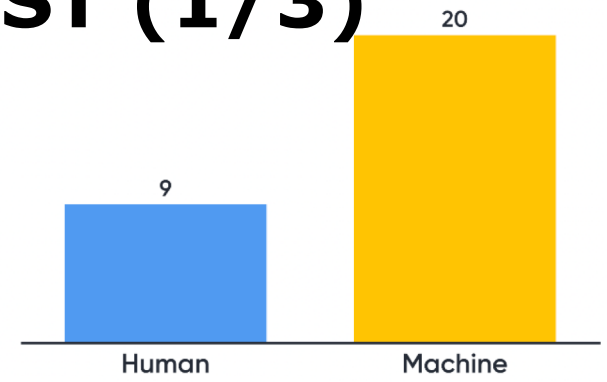
DEEP LEARNING

- Neural-based models have achieved superior results on various language-related tasks as compared to traditional machine learning models (e.g., SVM, logistic regression etc.).

TURING TEST

- Go to www.menti.com
- Code: 68 56 54

TURING TEST (1/3)



- Human or Machine?

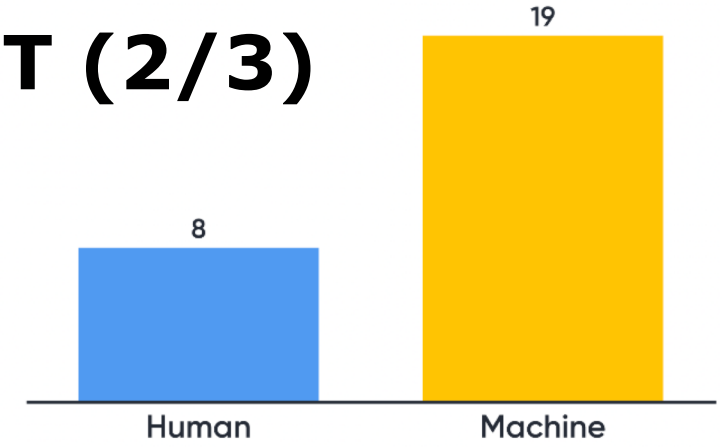
Input:

"The quick brown fox jumps over the lazy dog."

Output:

"The quick brown fox jumps over the lazy dog. This dog looks lazy! 'What's so great about that?' I asked, and the quick brown fox laughed, and the lazy dog licked his lips. 'I don't know. But you won't have to know, because once you're in… ' 'I know,' I said."

TURING TEST (2/3)



- Human or Machine?

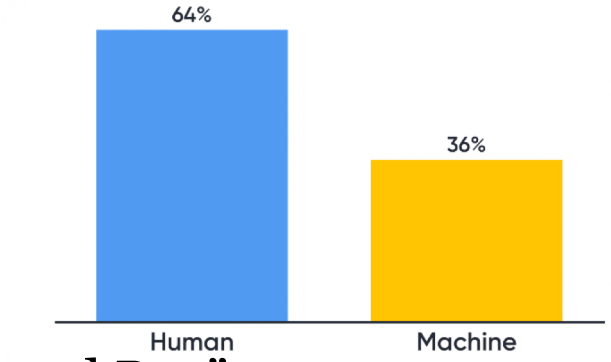
Input:

"John woke up this morning feeling very hungry."

Output:

"John woke up this morning feeling very hungry. This was due in part to the fact that while the restaurant is open through July, the rest of the business was closed for the Thanksgiving break. The meal started off with a salad of fresh vegetables, mixed with cheese and dressing. The potatoes had a nice brown finish and the salad was full of flavor."

TURING TEST (3/3)



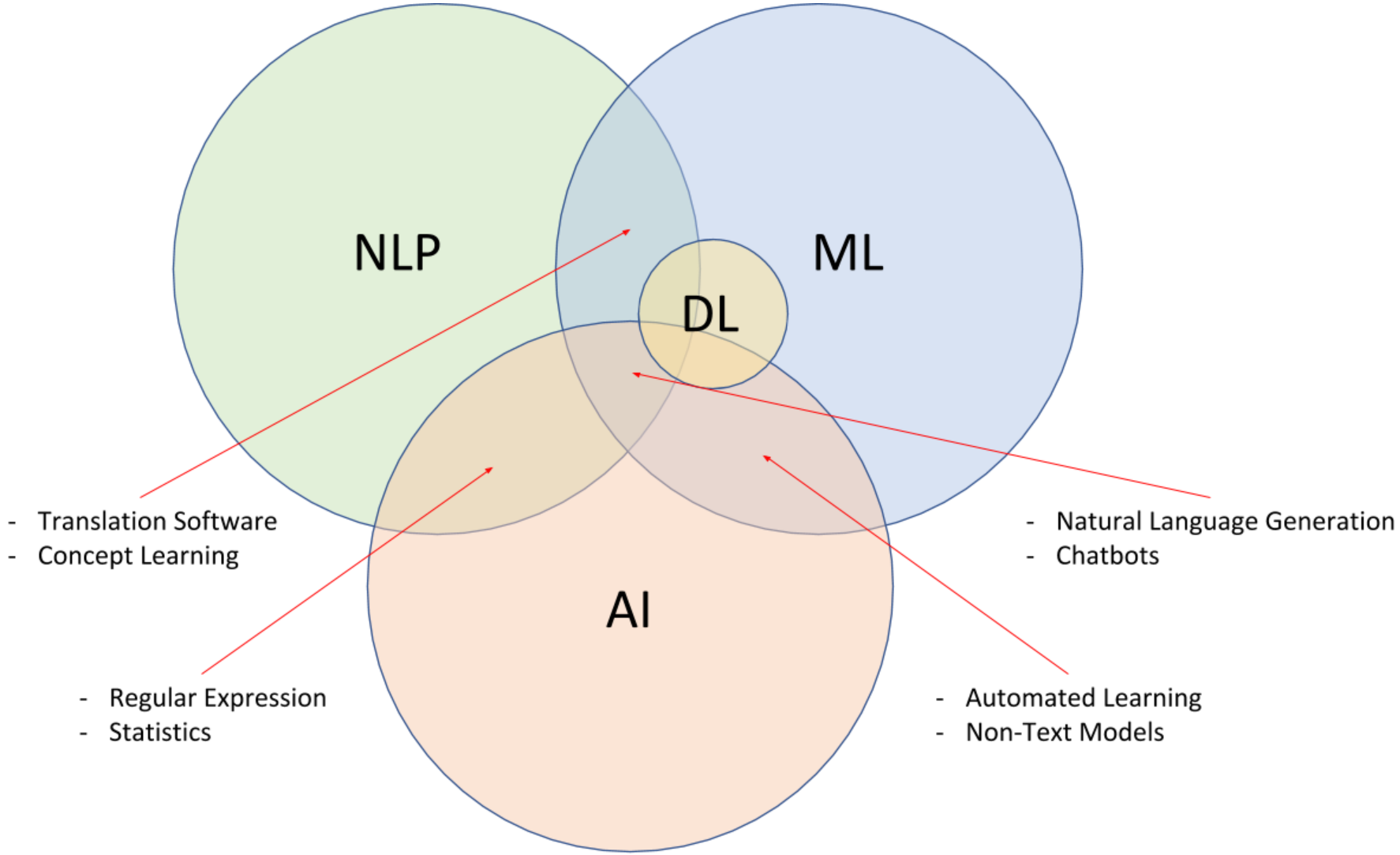
- Human or Machine?

Input:

"This is the story of a little dog named Boo"

Output:

"This is the story of a little dog named Boo. You see, Boo is a Labrador, a dog that loves people more than anything else. A little over a year ago, Boo became a part of The Dog Rescue of Arlington, Virginia. Boo was a puppy (now 2 years old) that had been abandoned. Boo's story was heartbreaking."


NLP
ML
DL
AI
- Translation Software
- Concept Learning
- Natural Language Generation
- Chatbots
- Regular Expression
- Statistics
- Automated Learning
- Non-Text Models

CONTENT

- Data Representation
- Deep Learning approachs on NLP

CONTENT

- Data Representation
- Deep Learning approachs on NLP

DATA REPRESENTATION

one hot representation

- Vocabulary: {man, woman, queen, king, Paris, Stockholm, apple, orange}

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
man	1	0	0	0	0	0	0	0
woman	0	1	0	0	0	0	0	0
queen	0	0	1	0	0	0	0	0
king	0	0	0	1	0	0	0	0
Paris	0	0	0	0	1	0	0	0
Stockholm	0	0	0	0	0	1	0	0
apple	0	0	0	0	0	0	1	0
orange	0	0	0		0	0	0	1

ONE HOT REPRESENTATION

- Each word is represented as a binary vector that is all zero values except the index of the its occurrence, which is marked with a 1.
- High-dimensional features

DISTRIBUTIONAL REPRESENTATION

- Distributional hypothesis: words appearing within similar context possess similar meaning.
- Word embbedding!
- Word embeddings are pre-trained on a task where the objective is to predict a word based on its context, typically using a shallow neural network.

WORD EMBEDDING

	man	woman	queen	king	Paris	Stockholm	apple	orange

WORD EMBEDDING

	man	woman	queen	king	Paris	Stockholm	apple	orange
gender	-1	1	-0.9	0.91	0.00	0.01	0.01	0.11

WORD EMBEDDING

	man	woman	queen	king	Paris	Stockholm	apple	orange
gender	-1	1	-0.9	0.91	0.00	0.01	0.01	0.11
royal	0.01	0.02	0.95	0.98	0.15	0.21	0.02	0.00

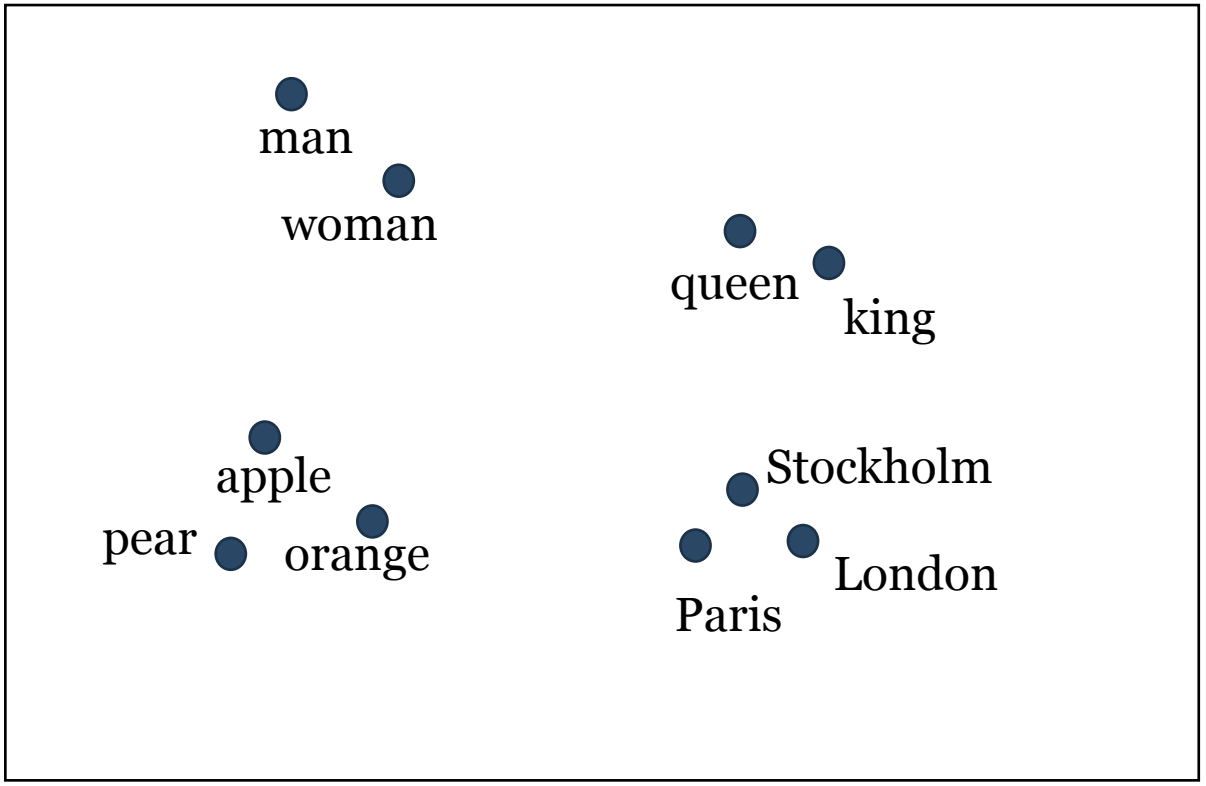
WORD EMBEDDING

	man	woman	queen	king	Paris	Stockholm	apple	orange
gender	-1	1	-0.9	0.91	0.00	0.01	0.01	0.11
royal	0.01	0.02	0.95	0.98	0.15	0.21	0.02	0.00
city	0.01	0.02	0.11	0.2	0.98	0.99	0.02	0.01

WORD EMBEDDING

	man	woman	queen	king	Paris	Stockholm	apple	orange
gender	-1	1	-0.9	0.91	0.00	0.01	0.01	0.11
royal	0.01	0.02	0.95	0.98	0.15	0.21	0.02	0.00
city	0.01	0.02	0.11	0.2	0.98	0.99	0.02	0.01
fruit	0.01	-0.01	0.01	0.01	0.01	0.00	0.98	0.99
.....								

VISUALIZE WORD EMBEDDING



WORD EMBEDDING

- Dense vector representations of words in lower dimensional space.
- Predict surrounding words of every word

CONTENT

- Data Representation
- Deep Learning approachs on NLP

SOME WORDS IN YOUR MIND?

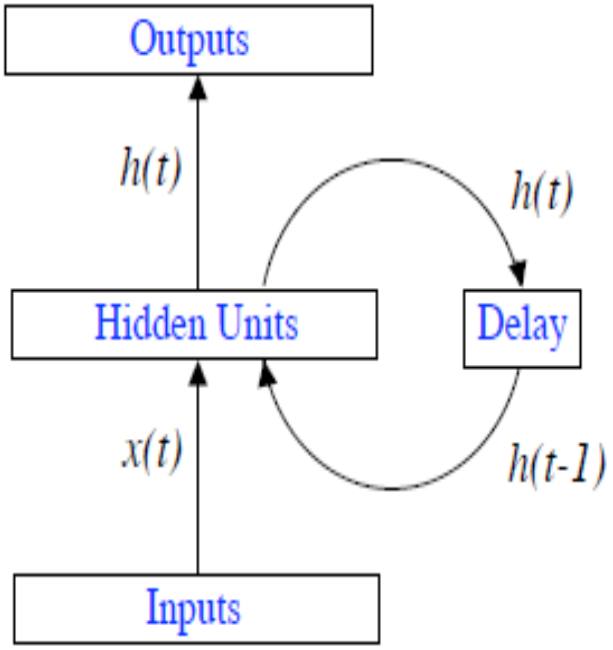
- RNN?
- LSTM?
- ?

RNN

- Recurrent Neural Networks
- Selectively pass information across sequence steps
- Allow a '**memory**' of previous inputs to persist in the network's internal state, and thereby influence the network output
- Time series prediction and sequence labelling tasks

WHAT ARE RNNS?

The simplest form of *fully recurrent neural network* is an MLP with the previous set of hidden unit activations feeding back into the network along with the inputs



$$h(t) = f_H(W_{IH}x(t) + W_{HH}h(t-1))$$

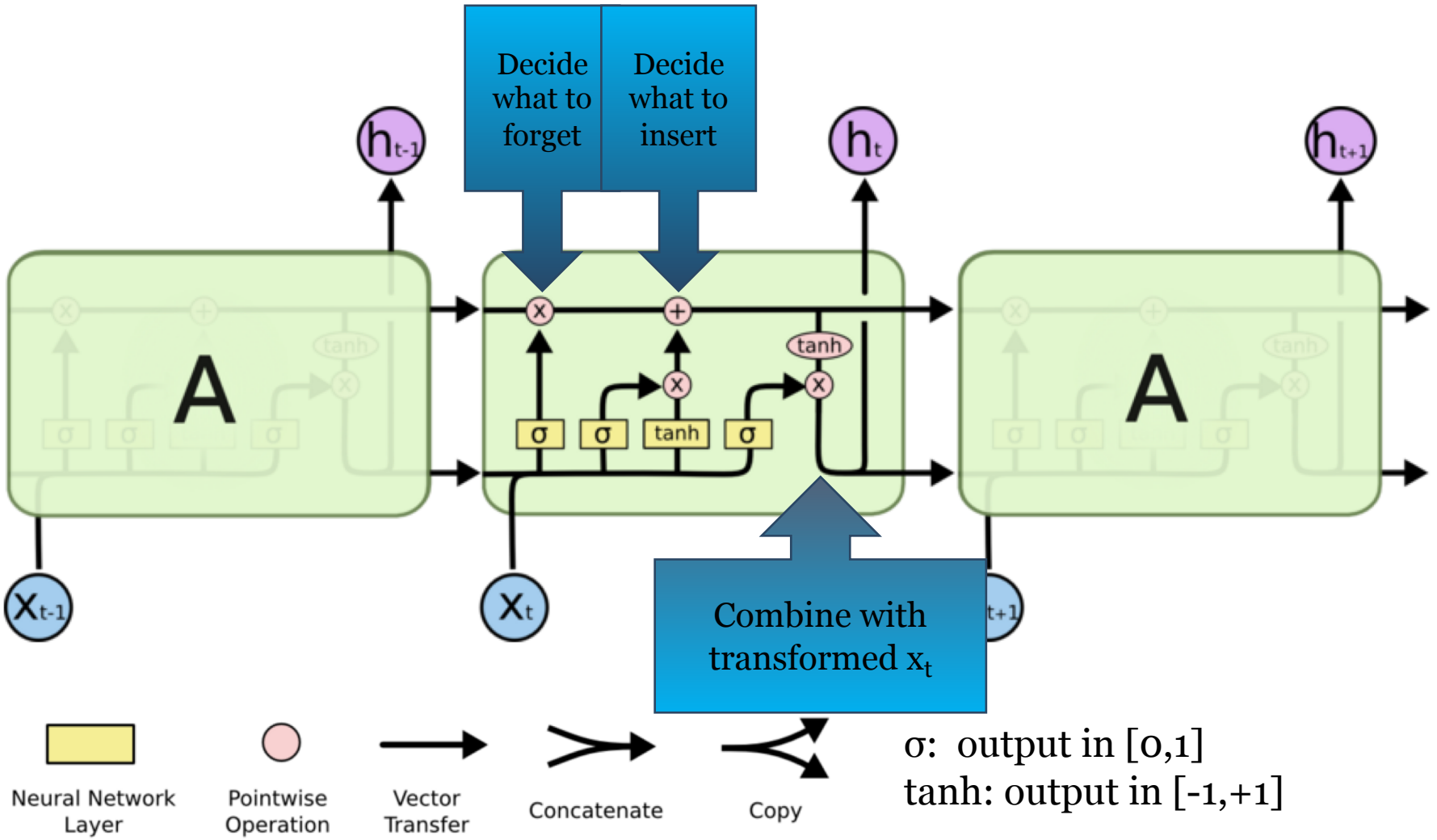
$$y(t) = f_O(W_{HO}h(t))$$

f_H and f_O are the activation function for hidden and output unit; W_{IH}, W_{HH}, and W_{HO} are connection weight matrices which are learnt by training.

RECURRENT NEURAL NETWORK

- Problems of RNN
  - Gradient vanishing
    - Cannot capture long dependencies
  - ”I grew up in China… I speak fluent *Chinese*.”
- Solutions
  - Long-short term memory (LSTM, Hochreiter, Schmidhuber, 1997)
  - Gated Recurrent Unit (GRU, Cho et al., 2014)

LSTM



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

CONTENT

- Data Representation
- Deep Learning approachs on NLP

ENCODER & DECODER

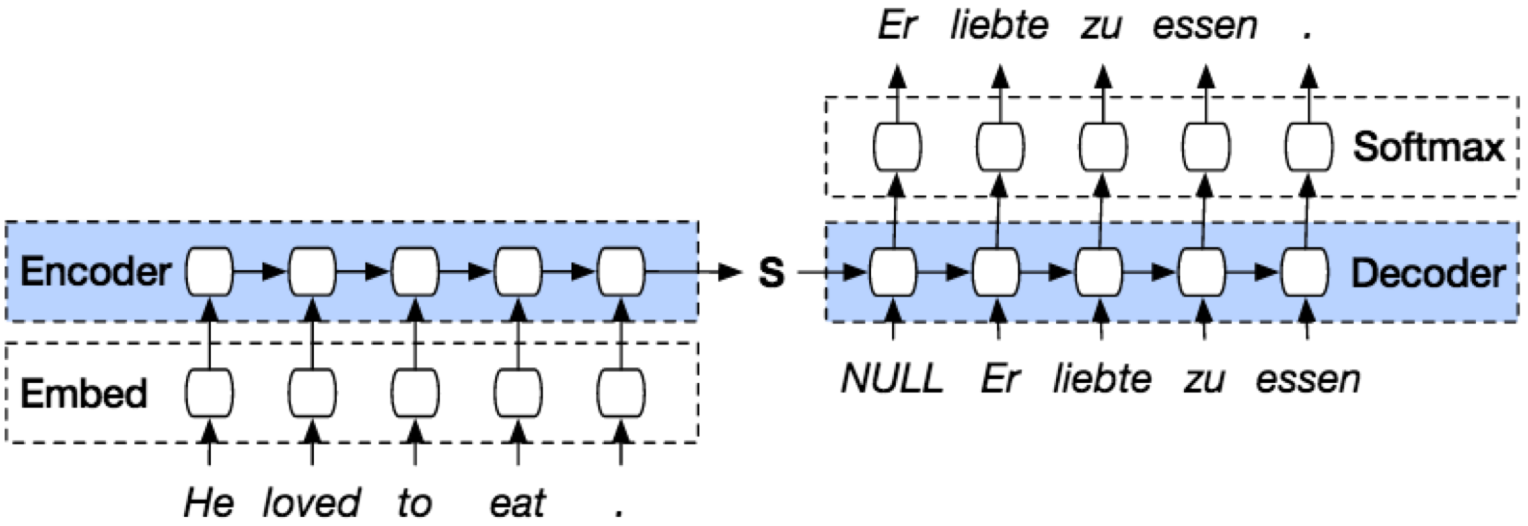


Image source: https://smerity.com/articles/2016/google_nmt_arch.html

ENCODER & DECODER

- Encoder: Maps input data to a different (often lower dimensional, compressed) feature representation (usually an RNN).
- Decoder: the model that takes the encoded representation and generates the output is called the Decoder

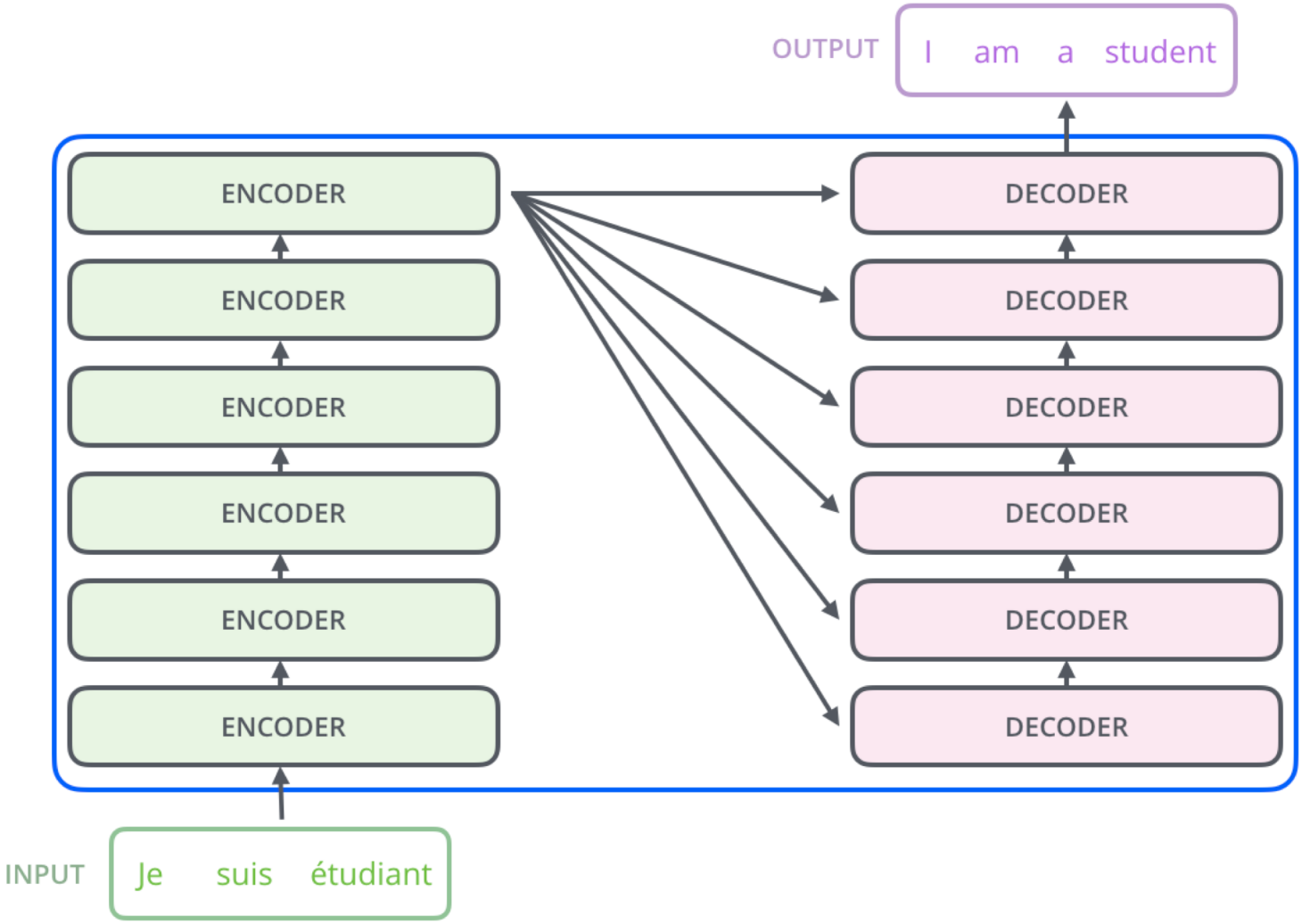
DEEP LEARNING MODELS (NLP)

- Transformer
- BERT
- Elmo
- ETNLP

TRANSFORMER

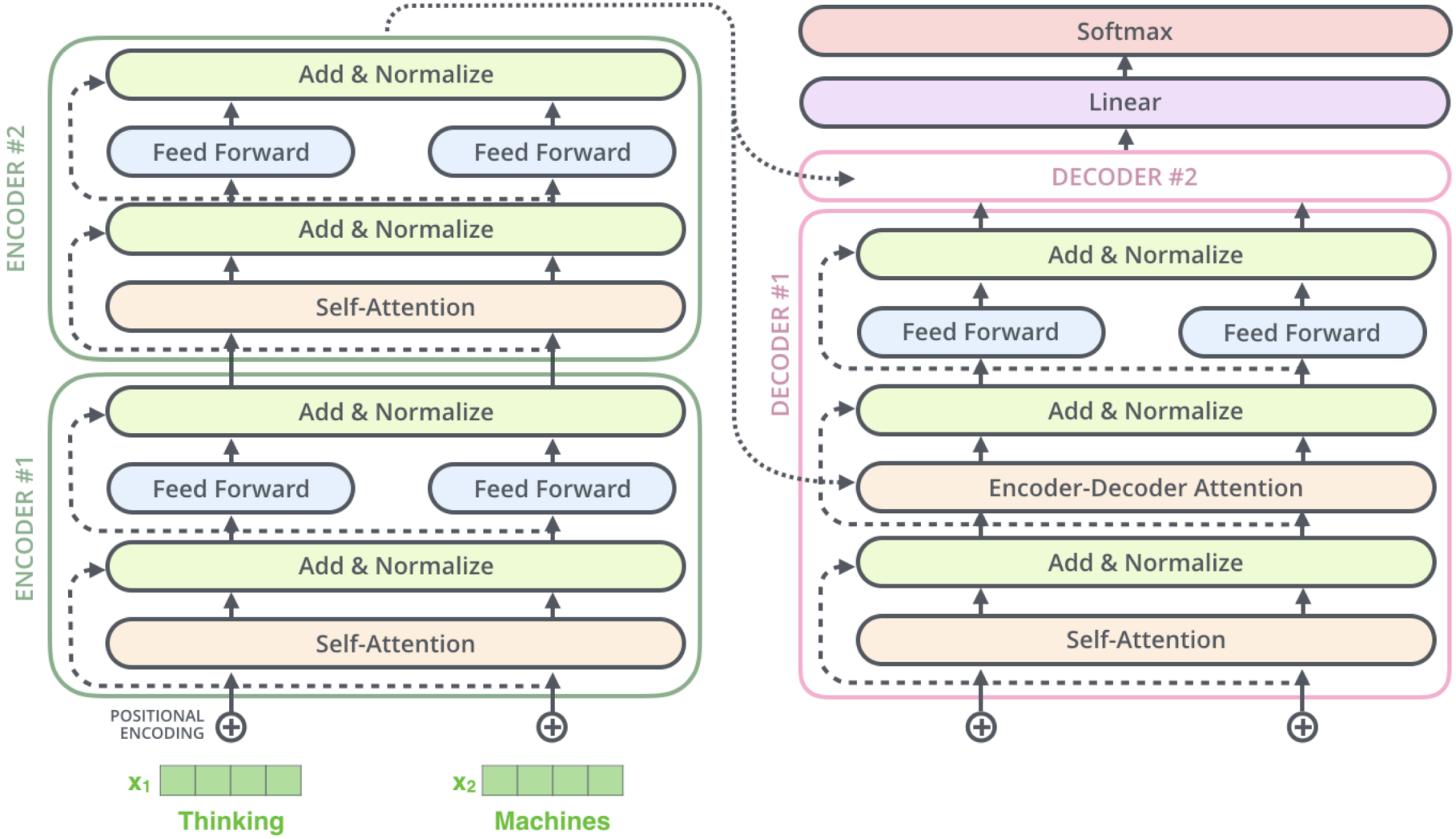
- From Google.
- A transformer is an encoder-decoder architecture model which uses **attention mechanisms** to forward a more complete picture of the whole sequence to the decoder at once rather than sequentially.

TRANSFORMER



http://jalammar.github.io/illustrated-transformer/

TRANSFORMER



http://jalammar.github.io/illustrated-transformer/

ADVANTAGE

- Effective modeling of long term dependencies among tokens in a temporal sequence.
- Eliminating the sequential dependency on previous tokens.

WELL SUITED PROBLEM

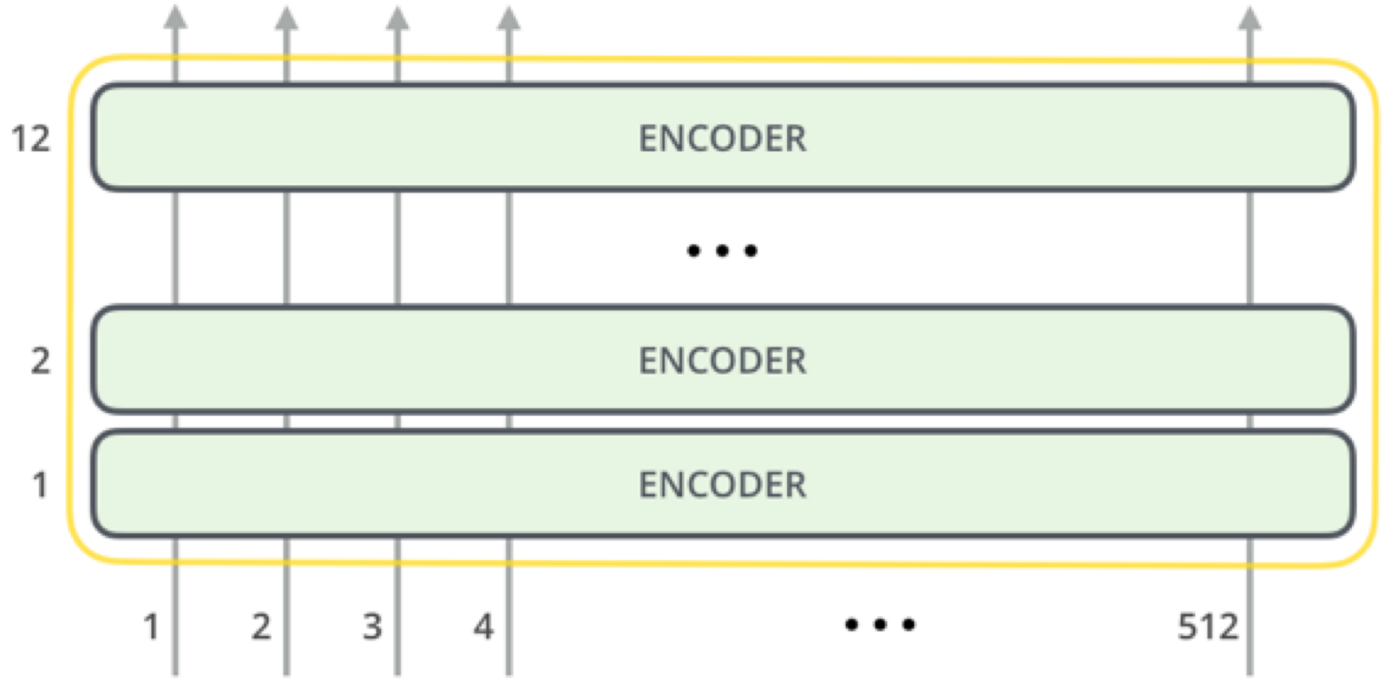
- Language translation
- Different inputs and outputs modalities, such as images and video
- co-reference resolution

BERT

- From Google
- Bidirectional Encoder Representations from Transformer (BERT)
- BERT brought everything together to build a bidirectional transformer-based language model using encoders rather than decoders.
- BERT is basically a trained Transformer Encoder stack.

[1] J.Devlin, M. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)

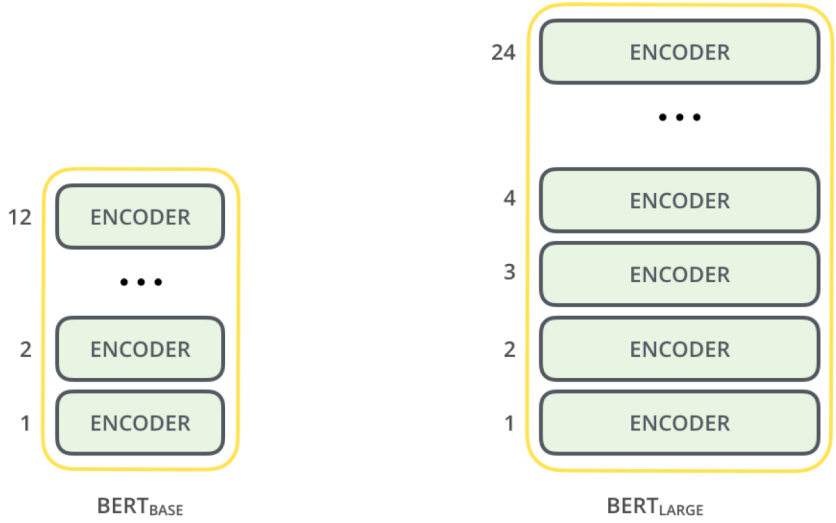
BERT



I am a student

TWO MAIN MODELS OF BERT

- The **BASE:** Number of transformer blocks (L): 12, Hidden layer size (H): 768 and Attention heads(A): 12
- The **LARGE:** Number of transformer blocks (L): 24, Hidden layer size (H): 1024 and Attention heads(A): 16



http://jalammar.github.io/illustrated-bert/

EXAMPLE OF USING BERT

- **Input**: the man went to the [MASK1] . he bought a [MASK2] of milk.
- **Labels**: [MASK1] = store; [MASK2] = liter

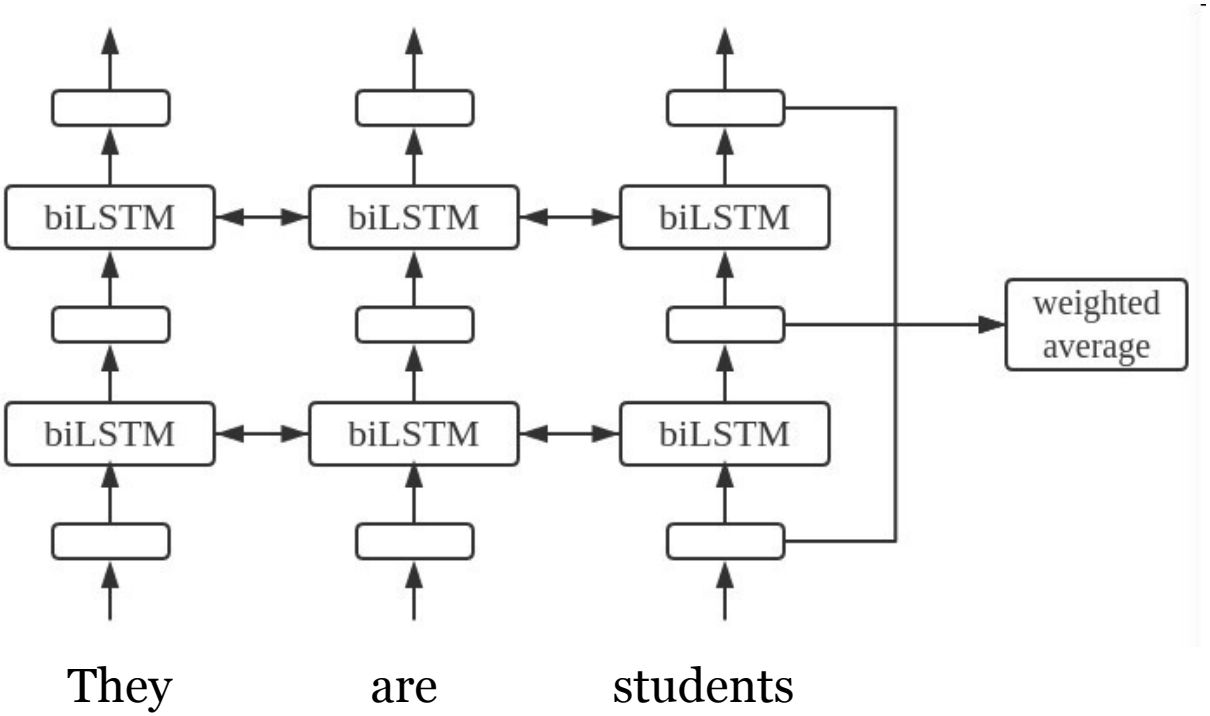
- **Sentence 1:** the man went to the store. **Sentence 2:** he bought a gallon of milk.
- Label: IsNextSentence

ELMO

- ELMo = Embeddings from Language Model
- ELMo uses a bidirectional LSTM and concatenates the left-to-right and right-to-left information (but not simultaneously).
- ELMo enables NLP models to better disambiguate between the correct sense of a given word.

ELMo: https://allennlp.org/elmo

ELMO


biLSTM
biLSTM
biLSTM
biLSTM
biLSTM
biLSTM
weighted average
They
are
students

BERT VS. ELMO

- ELMo and BERT
  - generate different word embeddings for a word
  - Context dependent

- ELMo uses LSTMS
- BERT uses Transformer - an attention based model with positional encodings to represent word positions
- ELMo learns word representation.
- BERT learns subword representation.

ETNLP

- A Toolkit for Extraction, Evaluation and Visualization of Pre-trained Word Embeddings
  - To **compare** quality of embedding models on the word analogy task.
  - To **extract** embedding vectors for other tasks.
  - To **visualize** the embedding space and compare between different embeddings.

https://github.com/vietnlp/etnlp

EXCERCISE 1 (30MIN)

- Play with RNN in tensorflow.

EXCERCISE 2 (60MIN)

- Play with ETNLP to feel the different output from different word embeddings.

Thanks

UMEÅ UNIVERSITY