



UPPSALA
UNIVERSITET

- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Machine learning, big data and artificial intelligence – Block 6

Måns Magnusson
Department of Statistics, Uppsala University

HT 2020



This week's lectures

- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
-
- Word embedding basics
 - Recurrent Neural Networks
 - Attention and Transformers
 - BERT



How do we represent words?

- One-hot encoding
 - A vector of length V (vocabulary size)

$\text{Uppsala} = [0, \dots, 1, \dots, 0] = 1_i$

- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

How do we represent words?

- One-hot encoding
 - A vector of length V (vocabulary size)

$$\text{Uppsala} = [0, \dots, 1, \dots, 0] = 1_i$$

- Word embeddings
 - A vector of length D (embedding dimension)

$$\text{Uppsala} = [-0.1231, \dots, 1.9001, \dots, 0.012]$$

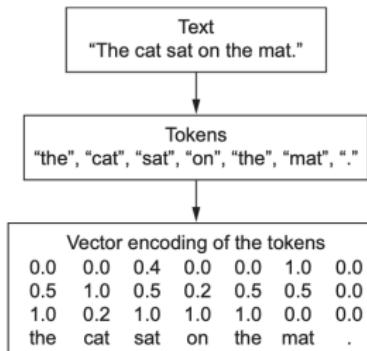


Figure: Representing words as word emnbeddings (Chollet and Allair, 2018, Fig. 6.1)

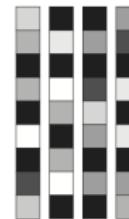


- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Word embeddings vs. One-Hot



One-hot word vectors:
- Sparse
- High-dimensional
- Hardcoded



Word embeddings:
- Dense
- Lower-dimensional
- Learned from data

Figure: One-Hot vs. Word embeddings (Chollet and Allair, 2018, Fig. 6.2)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Word embeddings

- A word type represent meaning in a low-dimensional semantic space Word embeddings
- The distributional hypothesis:
 - Harris (1954) and Firth (1957):
“A word is characterized by the company it keeps”
 - Semantics (broadly defined) is captured by context
- Lots of different embeddings:
word2vec, GloVe, Probabilistic Embeddings

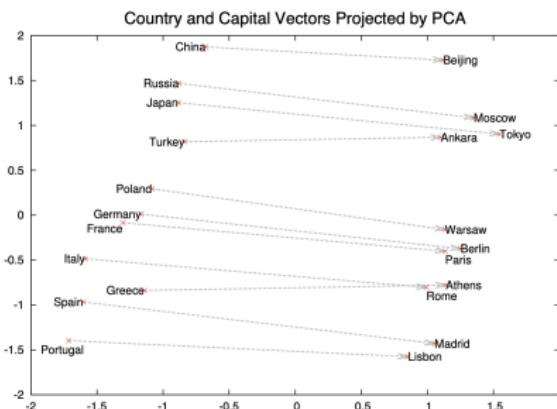


Figure: Word embedding properties (Mikolov et al, 2013)



Context Matters!

UPPSALA
UNIVERSITET

- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

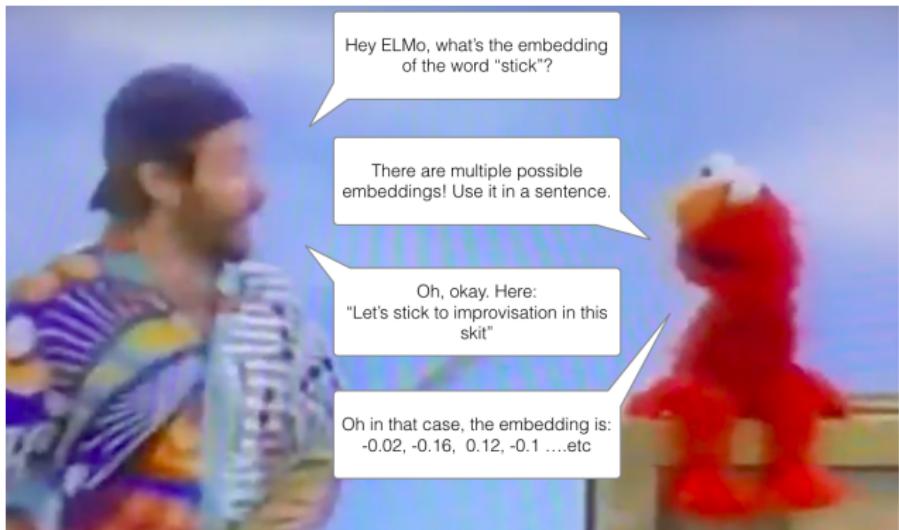


Figure: Context matters (Alammar, 2020)



Recurrent Neural Networks

- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT
- Recurrent Neural Networks, Recurrent Nets, RNN, ...
- Modeling of **temporal data structures**, such as
 - Time series data
 - Sequences of words (language models)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Recurrent Neural Networks

- Recurrent Neural Networks, Recurrent Nets, RNN, ...
- Modeling of **temporal data structures**, such as
 - Time series data
 - Sequences of words (language models)
- Examples of applications:
 - Text classification
 - Sequence / word classification
 - Time series predictions



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Recurrent Neural Networks

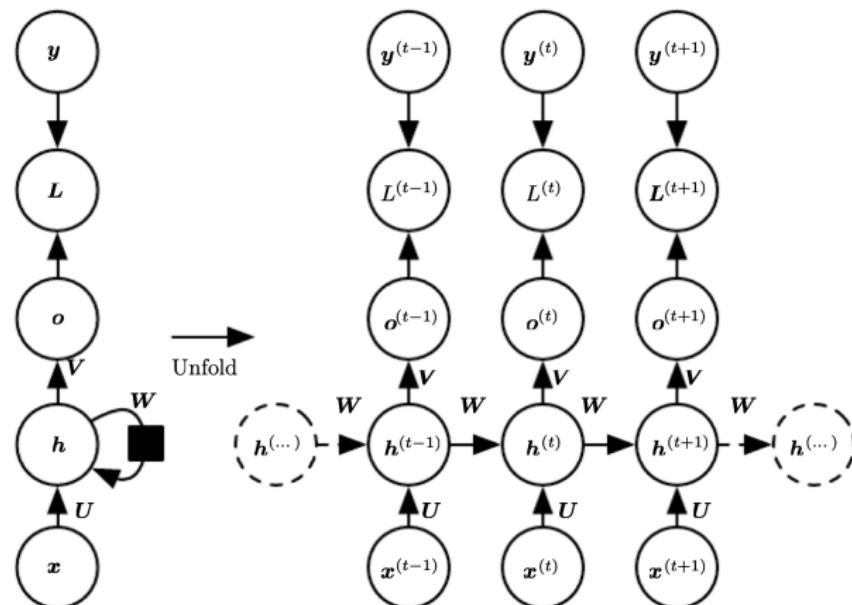


Figure: Recurrent Neural Network (Goodfellow et al, 2017, Fig. 10.3)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

$$a_t = b_t + Wh_{t-1} + Ux_t$$

$$h_t = \sigma a_t$$

$$o_t = c + Vh_t$$

$$\hat{y}_t = \text{softmax}(o_t)$$

Think of h_t as the "state" at timepoint t



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Recurrent network with one output

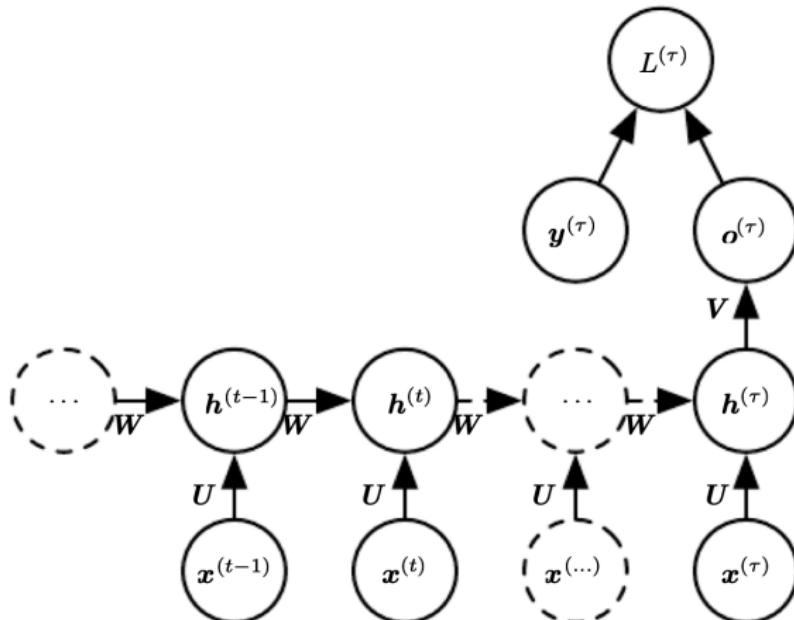


Figure: Recurrent Neural Network with one output (Goodfellow et al., 2017, Fig. 10.5)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Sequence to Sequence: Encoder-Decoder

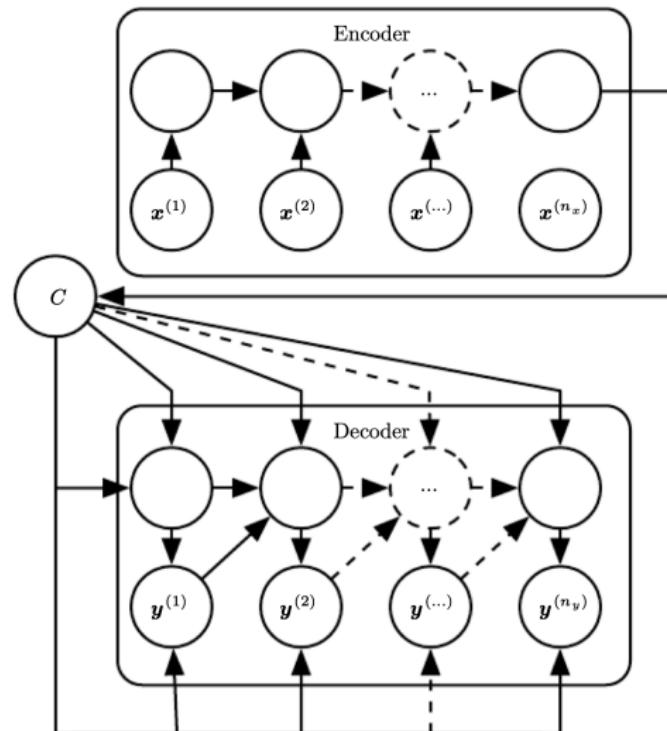


Figure: Encoder-Decoder Recurrent Networks (Goodfellow et al, 2017, Fig. 10.12)



Problems with RNN

- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
- Predicting sequences of different lengths
 - Exploding and vanishing gradients
 - Long-term dependencies





- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Bi-Directional RNN

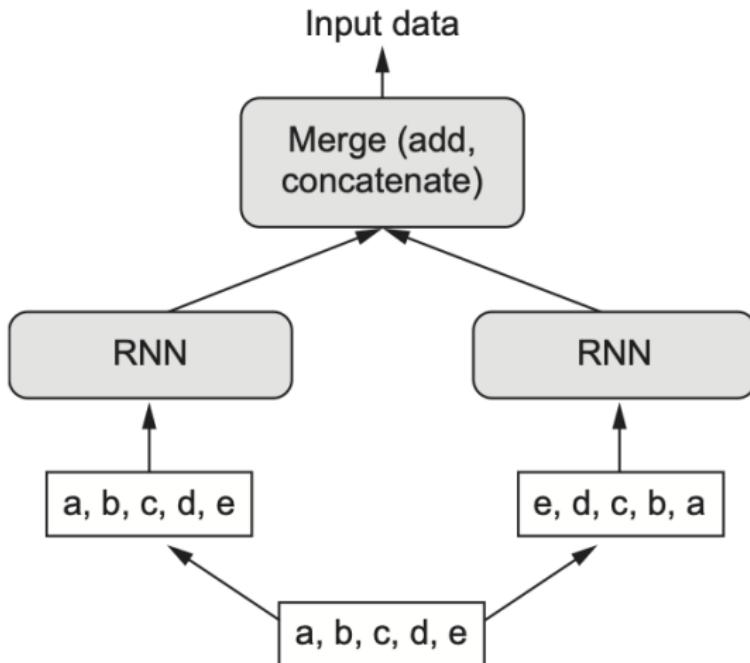


Figure: Bi-Directional RNN (Chollet and Allaire, 2018, Fig. 6.21)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

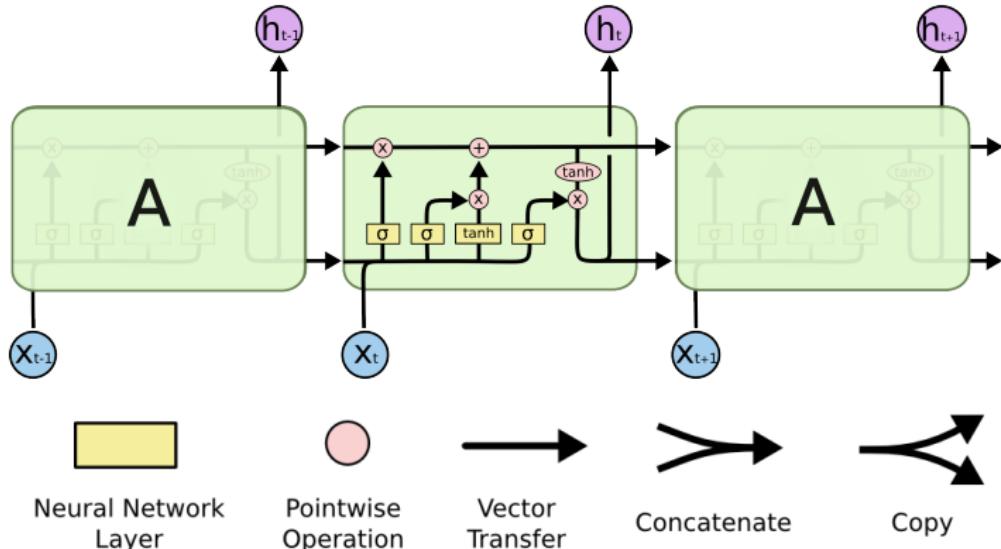


Figure: The LSTM (Olah, 2015)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

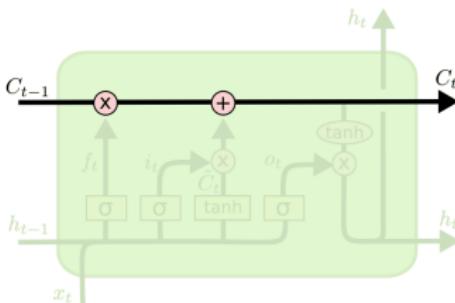
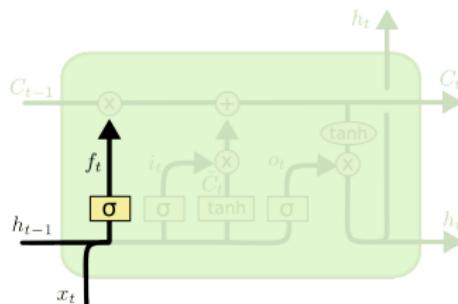


Figure: LSTM cell state, i.e. "carrybelt" (Olah, 2015)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

LSTM forget gate



$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

Figure: LSTM forget gate (Olah, 2015)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

LSTM input gate

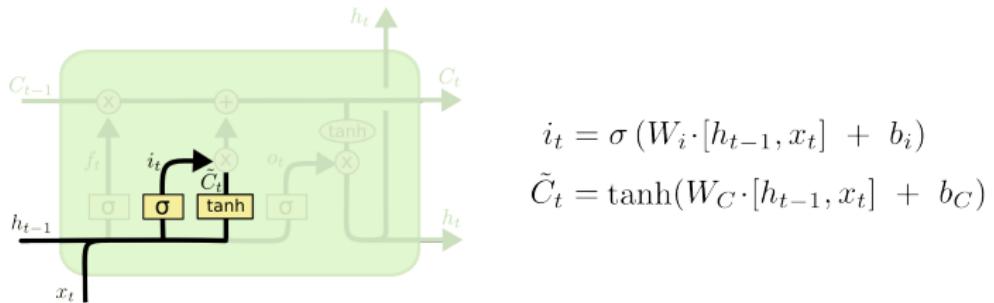


Figure: LSTM input gate (Olah, 2015)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

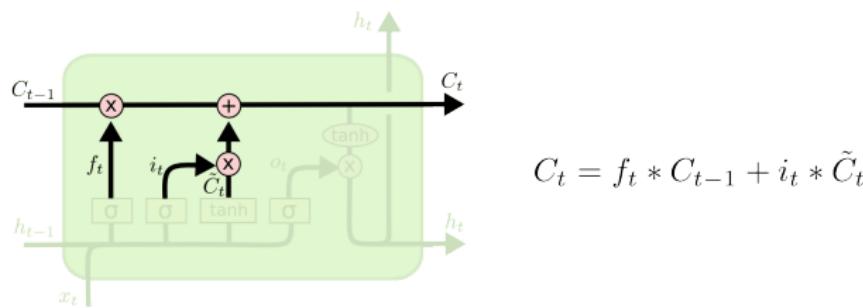
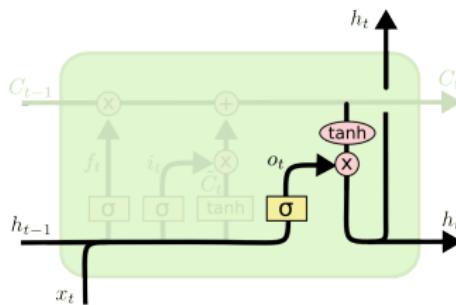


Figure: Update cell state (Olah, 2015)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

LSTM output gate



$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

Figure: LSTM output gate (Olah, 2015)



Problems

- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
- Still a **recurrent structure**,
(vanishing and exploding gradients)
 - Long-term dependencies still difficult
 - Hard to do **transfer learning**





- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT
- Introduced in 2017 in Vaswani et al. (2017)
- Behind the recent progress in NLP: BERT, GPT-2, GPT-3, etc.



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
-
- Introduced in 2017 in Vaswani et al. (2017)
 - Behind the recent progress in NLP: BERT, GPT-2, GPT-3, etc.
 - Becoming de-facto standard in industry and academia



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
-
- Introduced in 2017 in Vaswani et al. (2017)
 - Behind the recent progress in NLP: BERT, GPT-2, GPT-3, etc.
 - Becoming de-facto standard in industry and academia
 - Brings **transfer learning** to NLP



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
-
- Introduced in 2017 in Vaswani et al. (2017)
 - Behind the recent progress in NLP: BERT, GPT-2, GPT-3, etc.
 - Becoming de-facto standard in industry and academia
 - Brings **transfer learning** to NLP
 - Two large benefits:
 - Enables more **parallelism**
 - Better handling of **long-range dependencies**



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT



Figure: Attention (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Stacked Encoder-Decoder Structure

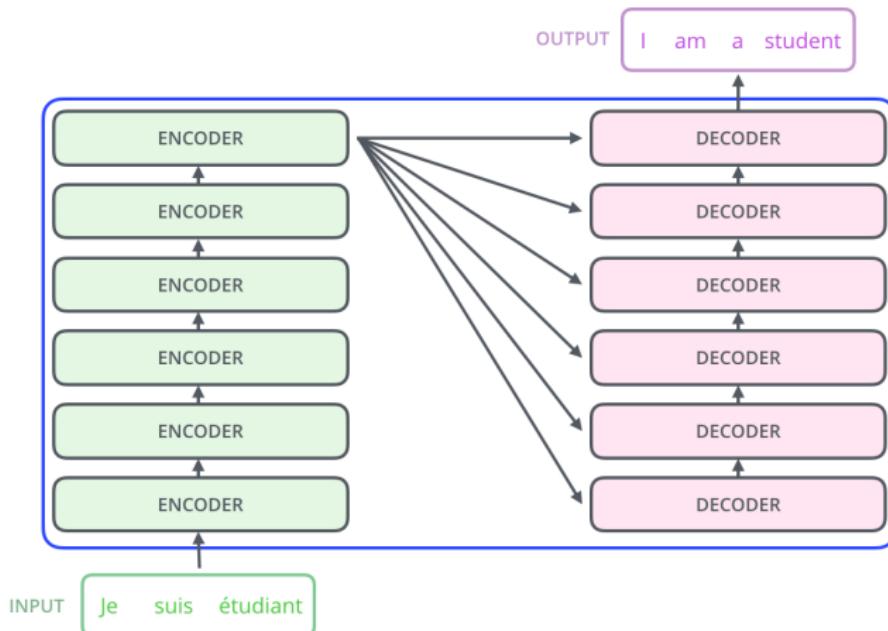


Figure: Attention (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Transformer

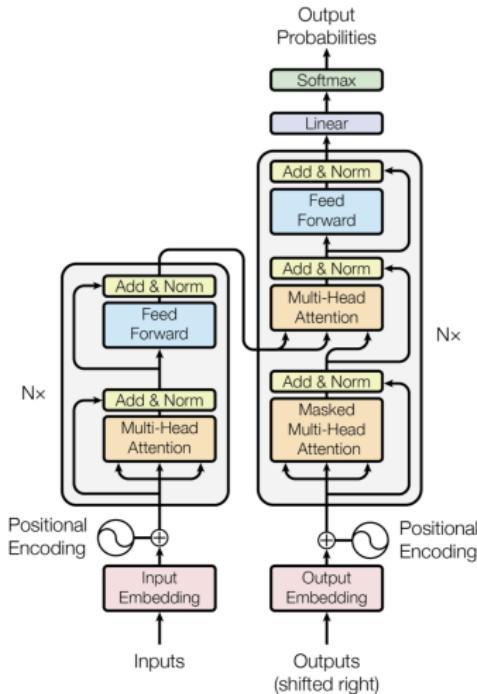


Figure: The Transformer Architecture (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

The encoder vs. the decoder

- Encoder:
 - Input: words (embeddings)
 - Output: contextualized embeddings
- Decoder:
 - Input: contextualized embeddings **and** previous words (embeddings)
 - Output: words (embeddings)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

The Encoder Layer

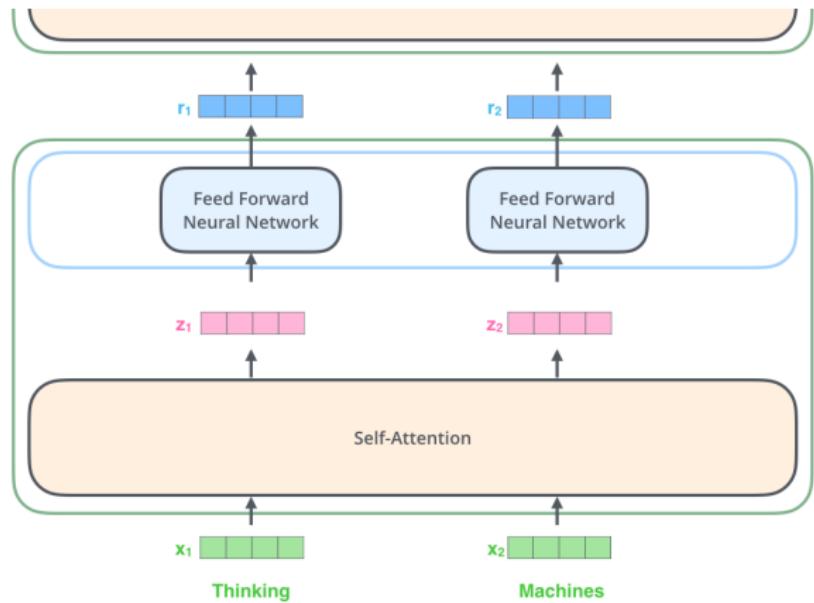


Figure: The Encoder Layer (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Scaled Dot-Product Attention

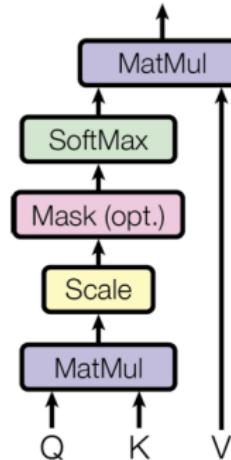


Figure: Scaled Dot-Product Attention (Vaswani et al., 2017)



The encoder vs. the decoder

- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT
-
- (Q)uery: Word i query other words
 - (K)ey: The other words return their key to i
 - (V)alue: The value of the other words to i





- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Computing Q, V and K

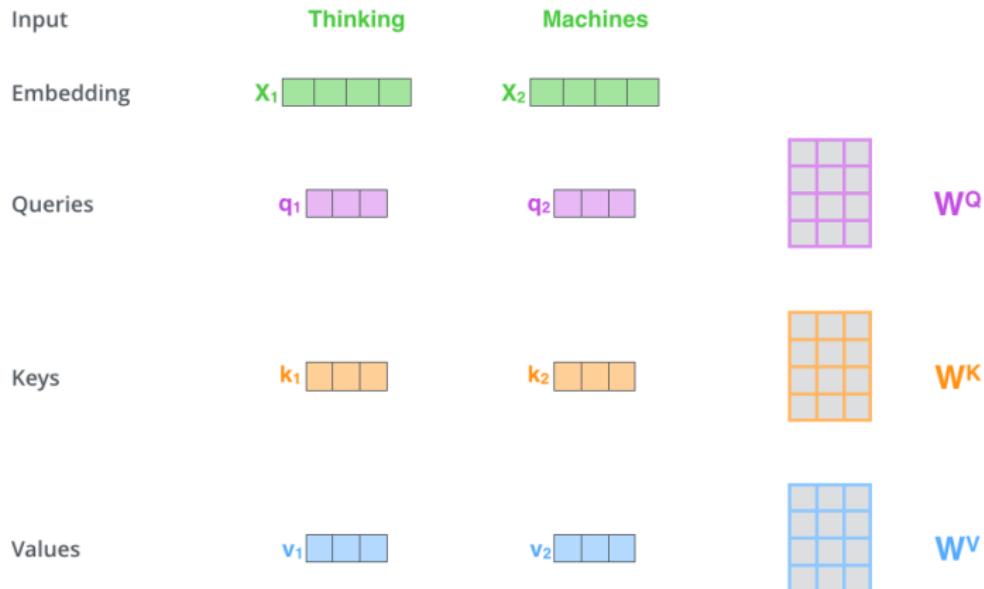


Figure: Attention heads (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Computing Self-Attention

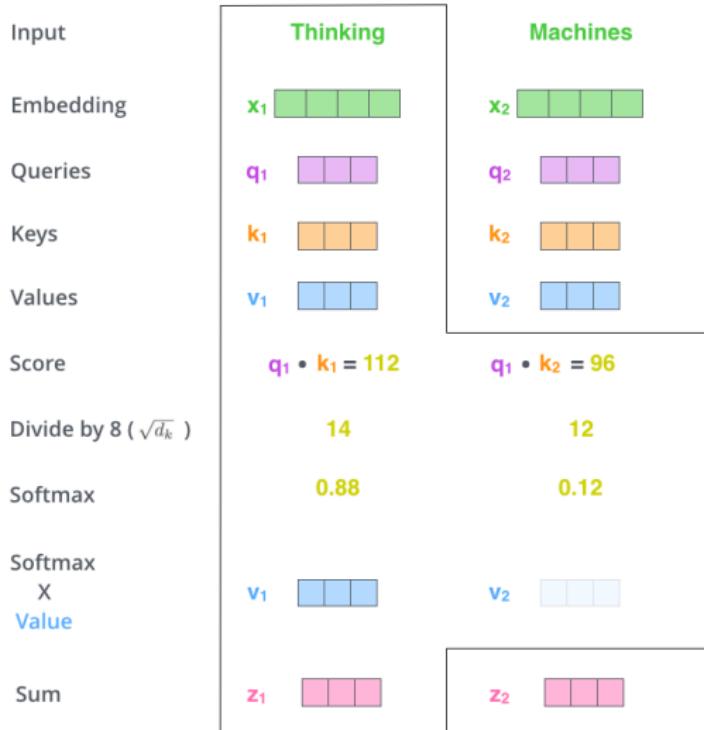


Figure: Attention (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Multi-Head Attention

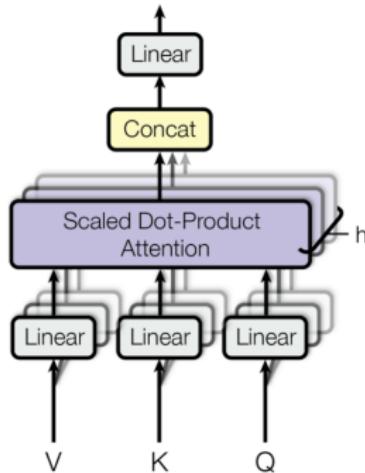


Figure: Scaled Dot-Product Attention (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Attentions Heads

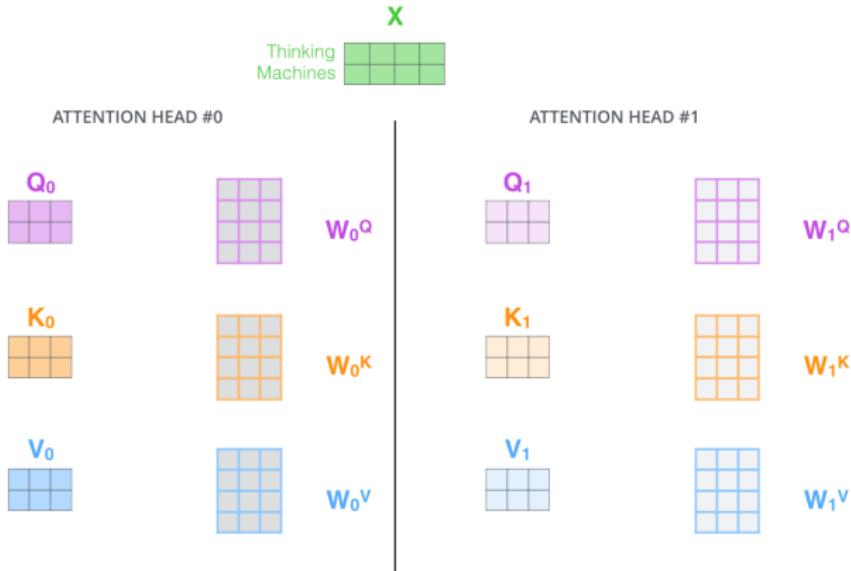


Figure: Attention heads (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Multi-head attention

- 1) This is our input sentence*
- 2) We embed each word*
- 3) Split into 8 heads. We multiply X or R with weight matrices
- 4) Calculate attention using the resulting $Q/K/V$ matrices
- 5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

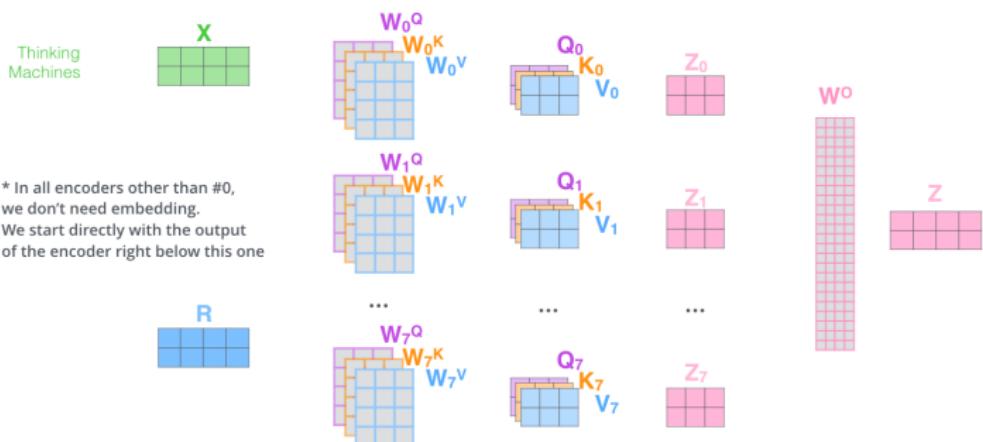


Figure: Attention heads (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Multi-Head Attention example

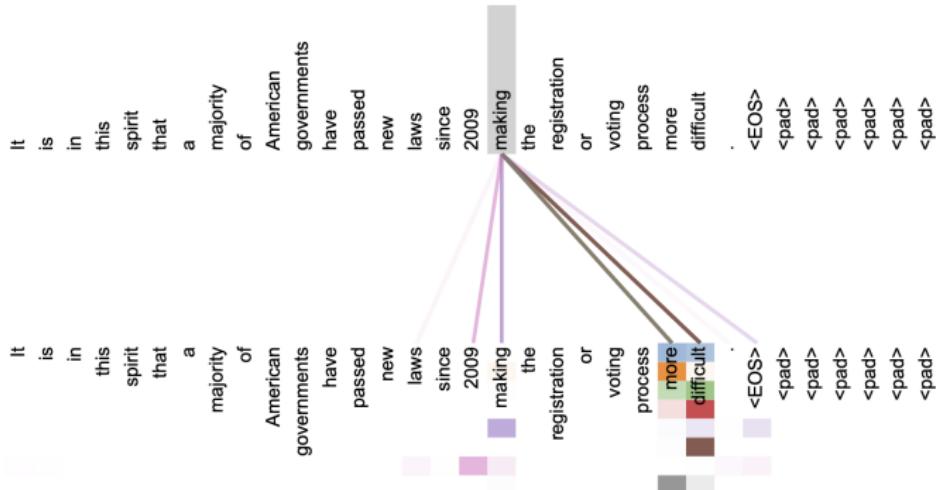


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb ‘making’, completing the phrase ‘making...more difficult’. Attentions here shown only for the word ‘making’. Different colors represent different heads. Best viewed in color.

Figure: Attention (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Positional Encoding

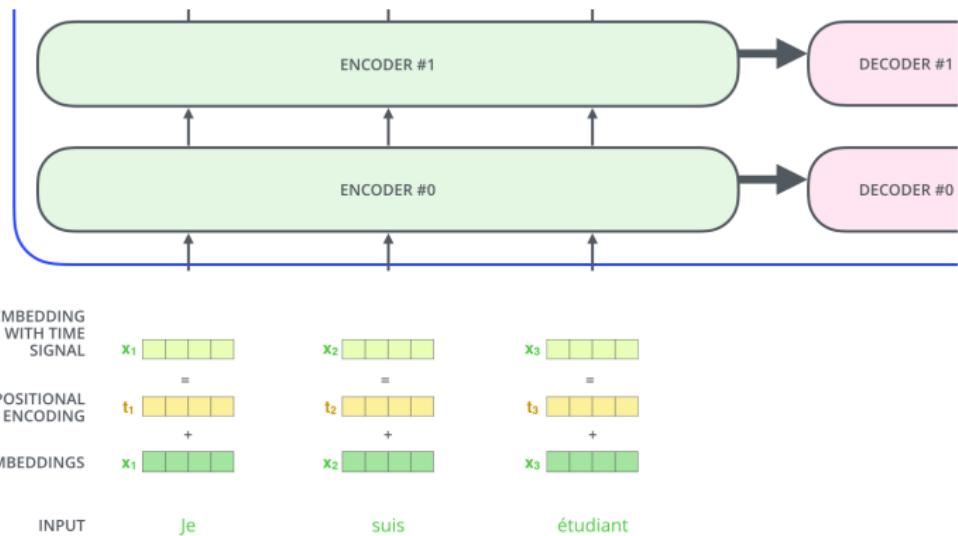


Figure: Attention heads (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Positional Encoding



Figure: Adding positional encodings to embeddings (Alammar, 2018)



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - BERT
 - Training BERT
 - Using BERT

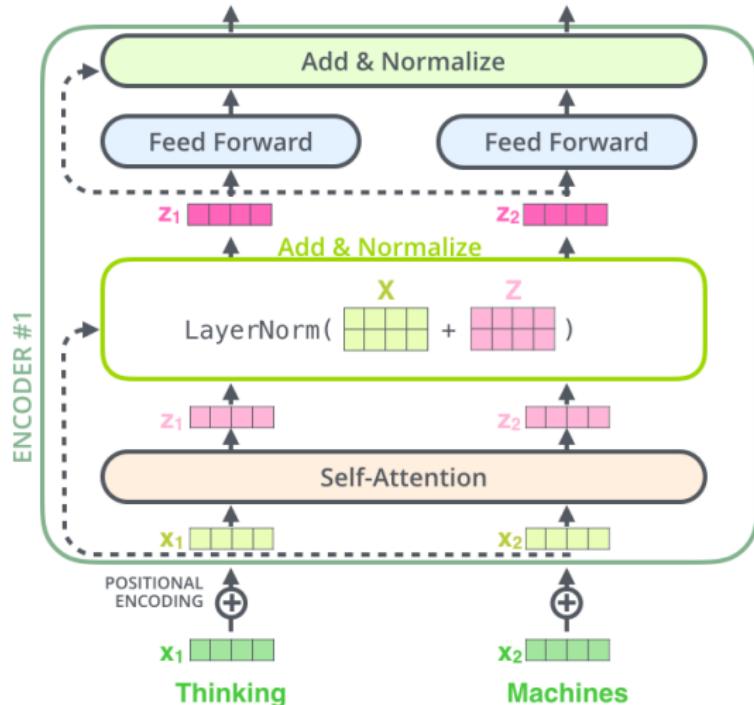


Figure: Add and Normalize (Alammar, 2018)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Transformer

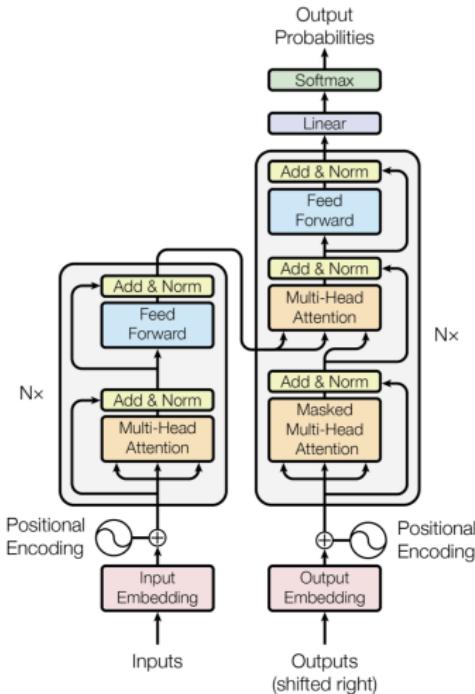


Figure: The Transformer Architecture (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- **BERT**
 - Training BERT
 - Using BERT
- Bidirectional Encoder Representations from Transformers
 - Introduced in 2018/2019 in Devlin et al. (2017)





- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - **BERT**
 - Training BERT
 - Using BERT
- Bidirectional Encoder Representations from Transformers
 - Introduced in 2018/2019 in Devlin et al. (2017)
 - **State-of-the-Art** in many text prediction tasks, such as
 - Question-Answering
 - Named-Entity Recognition
 - Text Classification



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - **BERT**
 - Training BERT
 - Using BERT
- Bidirectional Encoder Representations from Transformers
 - Introduced in 2018/2019 in Devlin et al. (2017)
 - **State-of-the-Art** in many text prediction tasks, such as
 - Question-Answering
 - Named-Entity Recognition
 - Text Classification
 - Becoming the de facto new standard in industry and NLP



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - **BERT**
 - Training BERT
 - Using BERT
- Bidirectional Encoder Representations from Transformers
 - Introduced in 2018/2019 in Devlin et al. (2017)
 - **State-of-the-Art** in many text prediction tasks, such as
 - Question-Answering
 - Named-Entity Recognition
 - Text Classification
 - Becoming the de facto new standard in industry and NLP
 - **Pre-trained** on a large number of books



- Word embeddings
 - Recurrent Neural Networks
 - LSTM
 - Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
 - **BERT**
 - Training BERT
 - Using BERT
- Bidirectional Encoder Representations from Transformers
 - Introduced in 2018/2019 in Devlin et al. (2017)
 - **State-of-the-Art** in many text prediction tasks, such as
 - Question-Answering
 - Named-Entity Recognition
 - Text Classification
 - Becoming the de facto new standard in industry and NLP
 - **Pre-trained** on a large number of books
 - Available both in English and Swedish (The National Library)

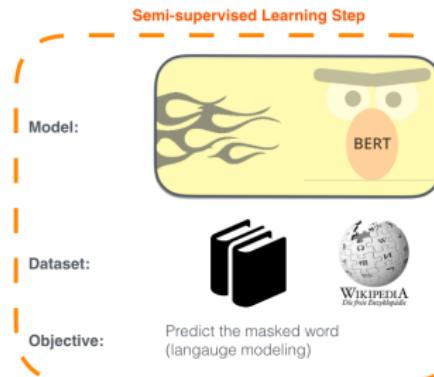


- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

BERT and transfer learning

1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.



2 - Supervised training on a specific task with a labeled dataset.

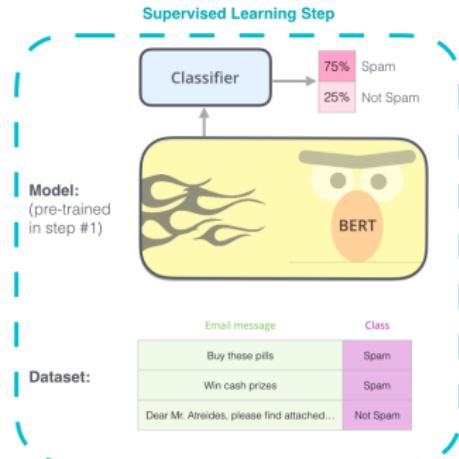


Figure: Using BERT for Transfer Learning (Alammar, 2019)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- **BERT**
 - Training BERT
 - Using BERT

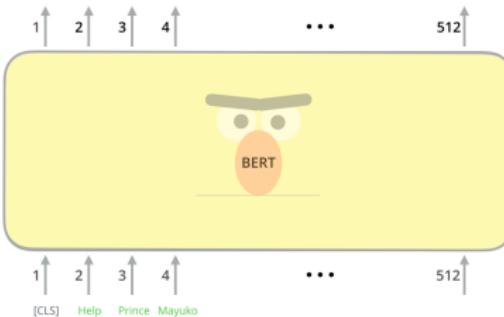


Figure: The BERT model (Alammar, 2019)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

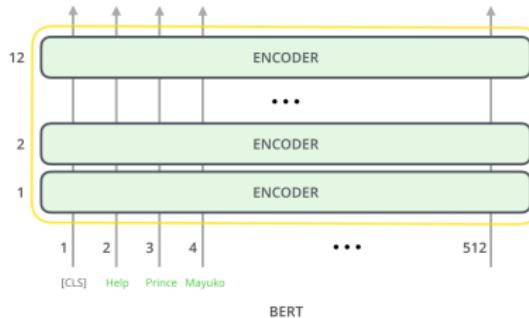


Figure: Opening up BERT (Vaswani et al., 2017)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Task 1: Masked Language Model

Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzyva

Randomly mask 15% of tokens

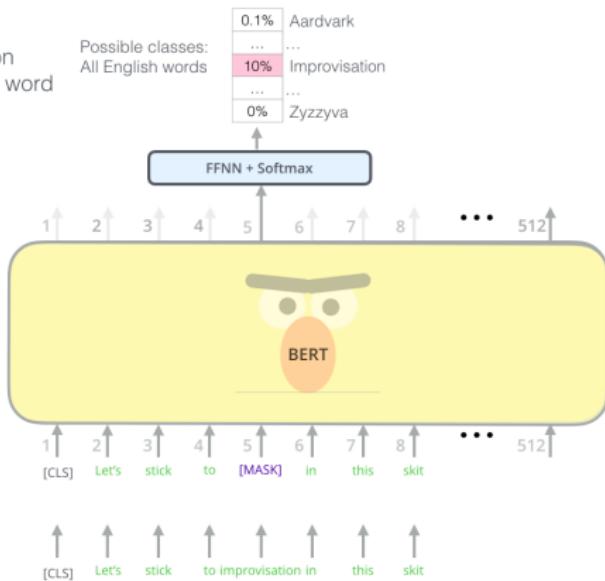


Figure: Masked Language Modeling (Alammar, 2019)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Next Sentence Prediction

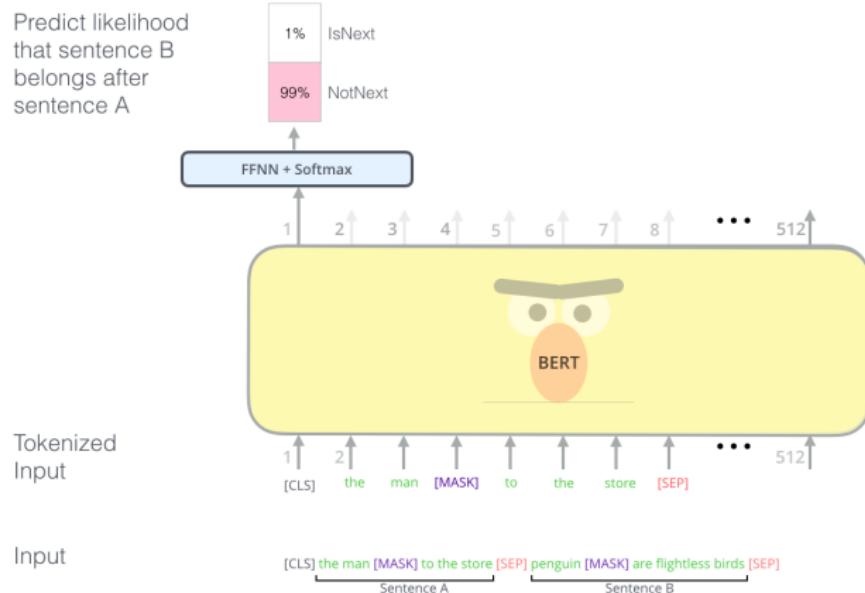


Figure: Next Sentence Prediction (Alammar, 2019)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

Using BERT for Classification

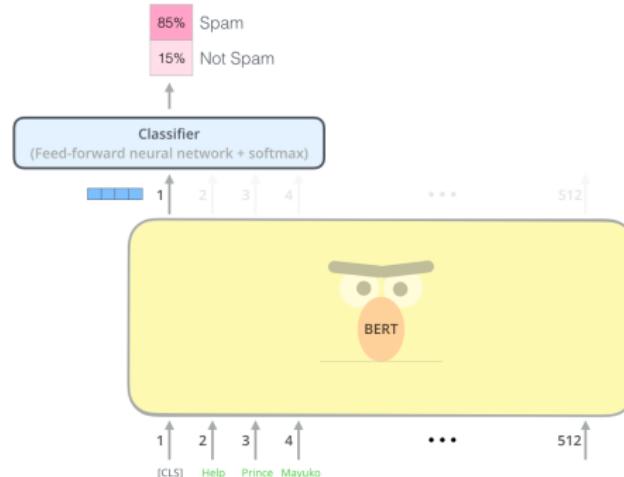


Figure: Using BERT for classification (Alammar, 2019)



- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

BERT and Contextualized embeddings

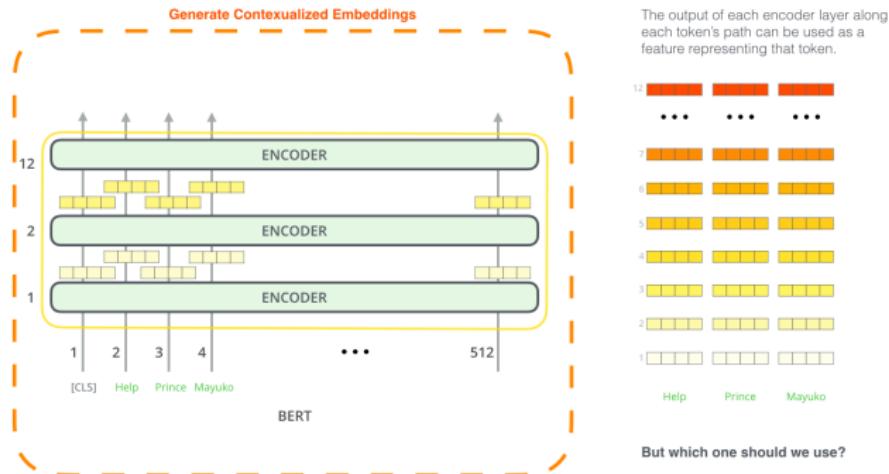


Figure: Contextualized Embeddings (Alammar, 2019)



Using Contextualized Embeddings

- Word embeddings
- Recurrent Neural Networks
 - LSTM
- Attention and Transformers
 - Attention
 - Multi-Head Attention
 - Positional encoding
 - Add and Normalize
- BERT
 - Training BERT
 - Using BERT

What is the best contextualized embedding for "Help" in that context?
For named-entity recognition task CoNLL-2003 NER

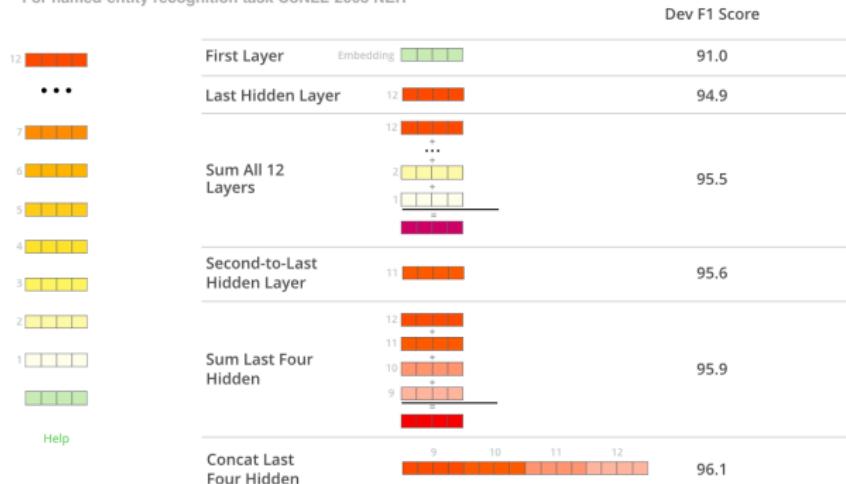


Figure: Using Contextualized Embeddings (Alammar, 2019)