UPPSALA UNIVERSITET

Machine learning – Block 5

Måns Magnusson
Department of Statistics, Uppsala University

Autumn 2022

UPPSALA UNIVERSITET

This week's lectures

- Word embedding basics
- Recurrent Neural Networks
- Attention and Transformers
- BERT

UPPSALA UNIVERSITET

Practicalities

- remember the project proposition deadline the 15th of December
- One lecture later this week on State of the Art in word embeddings and transformers (Väinö Yrjänäinen and Anders Östling)
- An additional guest lecture in january on fairness in AI and law (Holli Sargeant, Cambride University)

Assignment 3

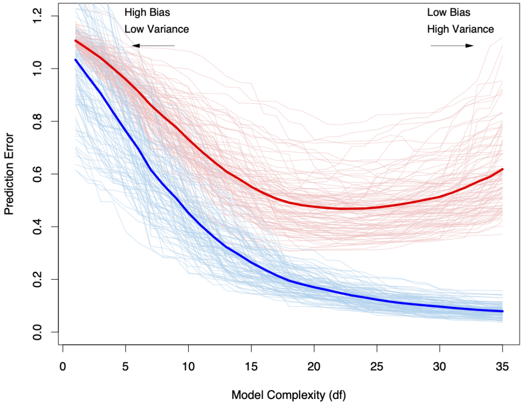
- remember the use of validation/test/train
- dont retrain your best model on test data
- clarify instructions: micro or macro precision and recall

Assignment 4: Evaluation

- Some further clarifications are needed
- It takes long time to run the models
- Better help in how to think when constructing networks

General principle

Figure: Test, training, and model complexity (Hastie et al, 2009, Figure 7)



Grid search vs. Random Search

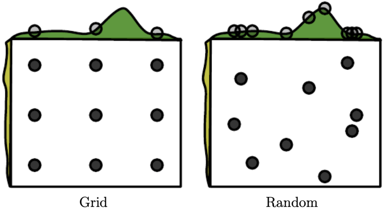


Figure: Grid search and random search (Goodfellow et al, 2017, Fig. 11.2)

How do we represent words?

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

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

How do we represent words?

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

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

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

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

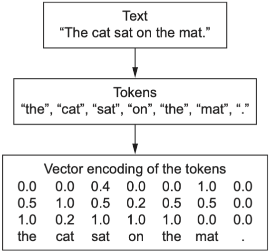


Figure: Representing words as word emnbeddings (Chollet and Allair, 2018, Fig. 6.1)

Word embeddings vs. One-Hot

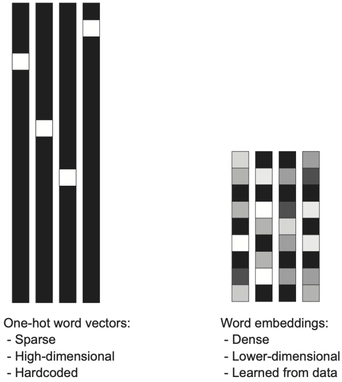


Figure: One-Hot vs. Word embeddings (Chollet and Allair, 2018, Fig. 6.2)

UPPSALA UNIVERSITET

Word embeddings

The quick brown fox jumps over the lazy dog.

- A word type represent meaning in a low-dimensional semantic space
- 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

Word embeddings

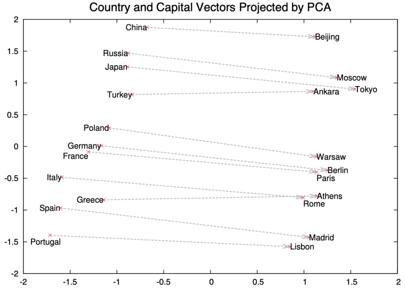


Figure: Word embedding properties (Mikolov et al, 2013)

king - man + woman $\approx$ queen

Context Matters!

Figure: Context matters (Alammar, 2020)

UPPSALA UNIVERSITET

Recurrent Neural Networks

- Recurrent Neural Networks, Recurrent Nets, RNN, ...
- Modeling of temporal data structures, such as
  - Time series data
  - Sequences of words (language models)

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

Recurrent Neural Networks

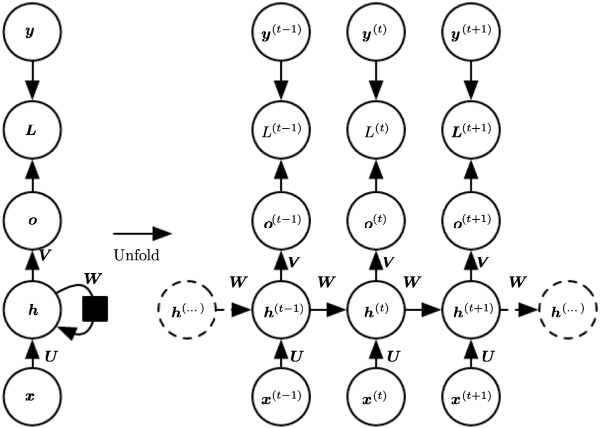


Figure: Recurrent Neural Network (Goodfellow et al, 2017, Fig. 10.3)

UPPSALA UNIVERSITET

Recurrent Neural Networks

$$
\begin{aligned}
a_t &= b + Wh_{t-1} + Ux_t \\
h_t &= \sigma_1(a_t) \\
o_t &= c + Vh_t \\
\hat{y}_t &= \sigma_{\text{output}}(o_t) = \text{softmax}(o_t)
\end{aligned}
$$

Think of h_t as the "state" at timepoint t

Recurrent network with one output

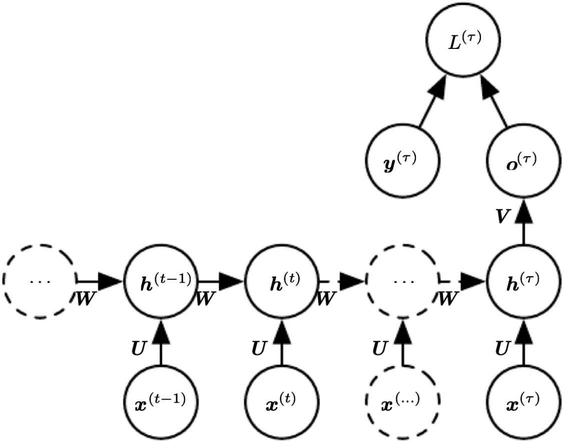


Figure: Recurrent Neural Network with one output (Goodfellow et al, 2017, Fig. 10.5)

Sequence to Sequence: Encoder-Decoder

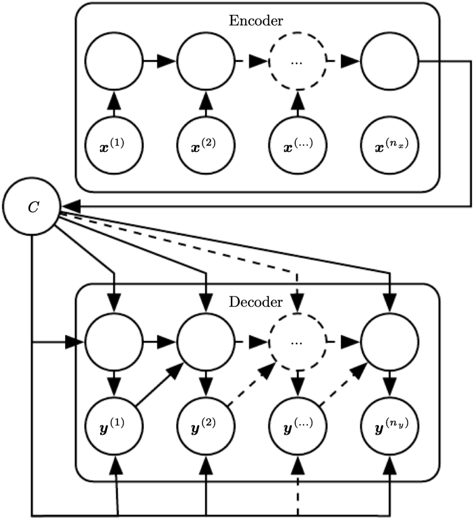


Figure: Encoder-Decoder Recurrent Networks (Goodfellow et al, 2017, Fig. 10.12)

Problems with RNN

- Predicting sequences of different lengths
- Exploding and vanishing gradients
- Long-term dependencies

Bi-Directional RNN

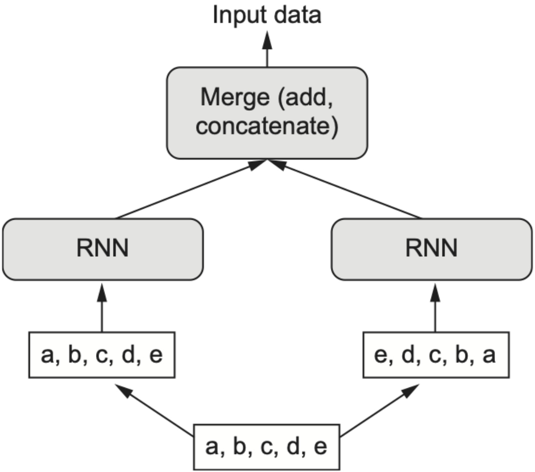


Figure: Bi-Directional RNN (Chollet and Allaire, 2018, Fig. 6.21)

UPPSALA UNIVERSITET

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

Long Short-Term Memory (LSTM)

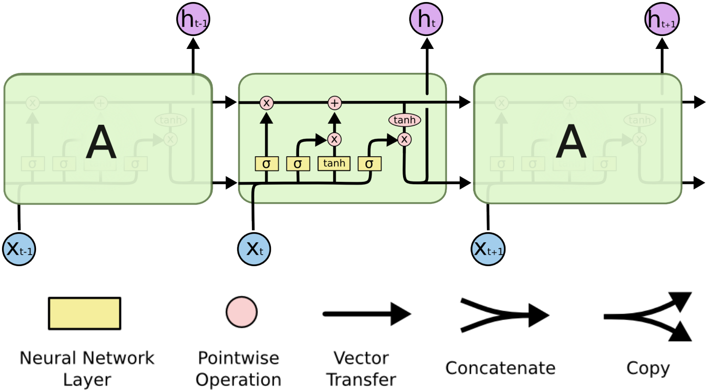


Figure: The LSTM (Olah, 2015)

LSTM cell state

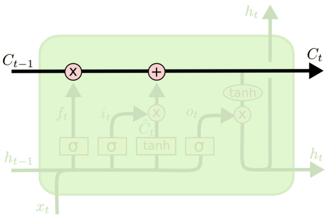
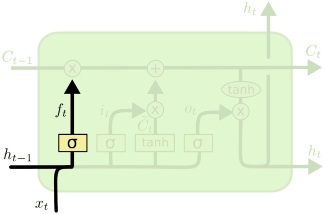


Figure: LSTM cell state, i.e. "carrybelt" (Olah, 2015)

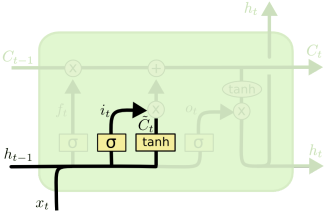
LSTM forget gate



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

Figure: LSTM forget gate (Olah, 2015)

LSTM input gate

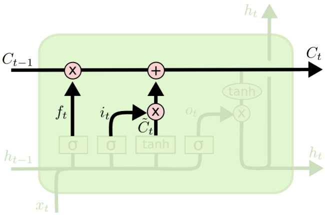


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Figure: LSTM input gate (Olah, 2015)

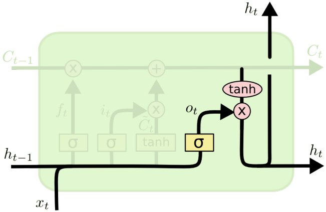
LSTM cell state update



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure: Update cell state (Olah, 2015)

LSTM output gate



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

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

Figure: LSTM output gate (Olah, 2015)

UPPSALA UNIVERSITET

Problems

- Still a recurrent structure,
  (vanishing and exploding gradients)
- Long-term dependencies still difficult
- Hard to do transfer learning

Transformer

- Introduced in 2017 in Vaswani et al. (2017)
- Behind the recent progress in NLP: BERT, GPT-2, GPT-3, etc.

Transformer

- 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

Transformer

- 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

Transformer

- 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

A Sequence-to-Sequence Model

Figure: Attention (Vaswani et al., 2017)

Stacked Encoder-Decoder Structure

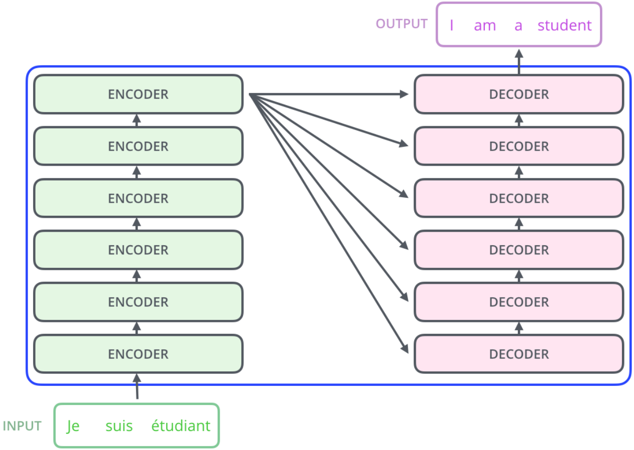


Figure: Attention (Vaswani et al., 2017)

Transformer

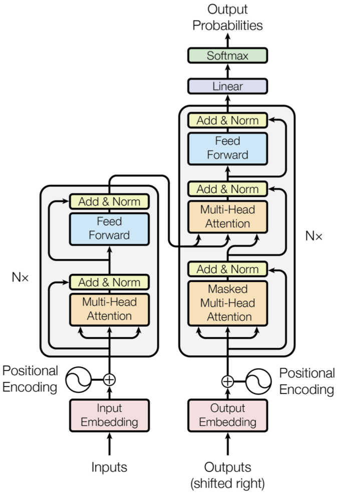


Figure: The Transformer Architecture (Vaswani et al., 2017)

The encoder vs. the decoder

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

UPPSALA UNIVERSITET

The Encoder Layer

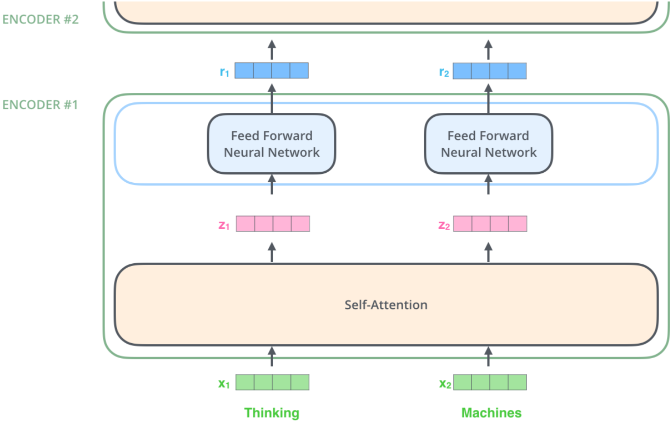


Figure: The Encoder Layer (Alammar, 2018)

Scaled Dot-Product Attention

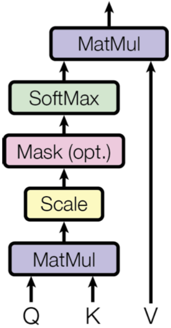


Figure: Scaled Dot-Product Attention (Vaswani et al., 2017)

UPPSALA UNIVERSITET

The encoder vs. the decoder

- (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

Computing Q, V and K

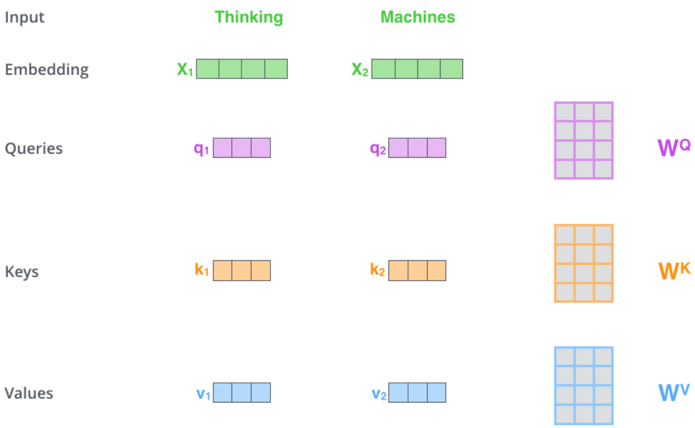


Figure: Attention heads (Alammar, 2018)

Computing Self-Attention

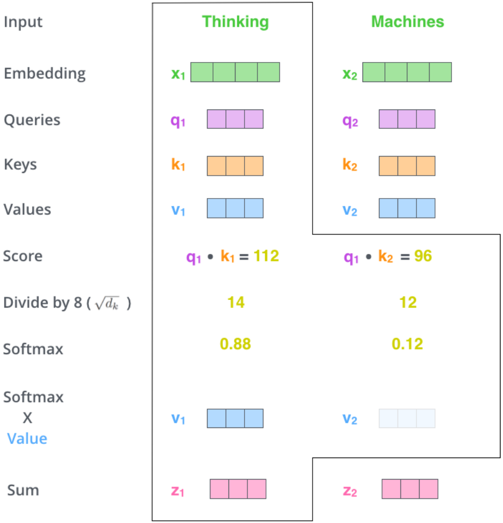


Figure: Attention (Alammar, 2018)

Multi-Head Attention

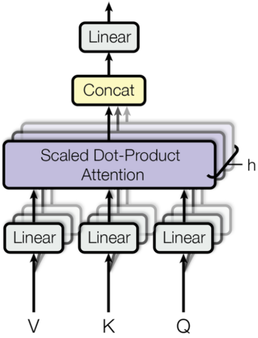


Figure: Scaled Dot-Product Attention (Vaswani et al., 2017)

Attentions Heads

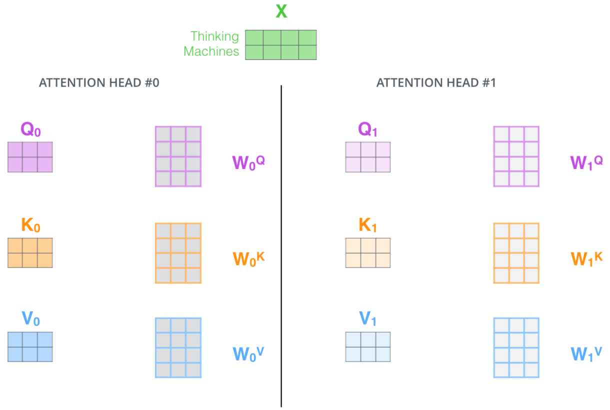


Figure: Attention heads (Alammar, 2018)

Multi-head attention

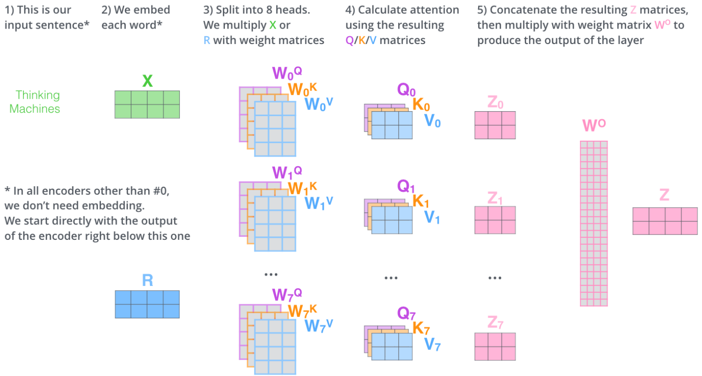


Figure: Attention heads (Alammar, 2018)

Multi-Head Attention example

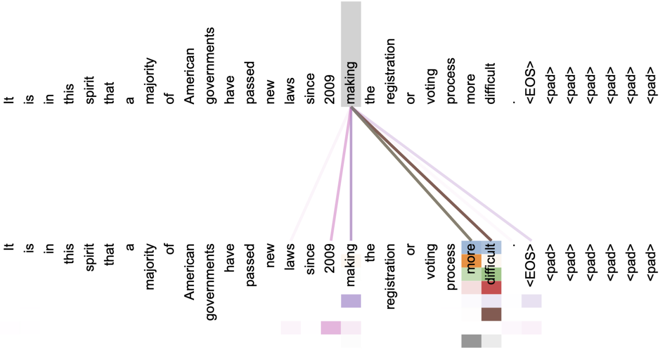


Figure 3: An example of the attention mechanism following long-distance dependencies in the encoder self-attention in layer 5 of 6. Many of the attention heads attend to a distant dependency of the verb 'making', completing the phrase 'making...more difficult'. Attentions here shown only for the word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

Figure: Attention (Vaswani et al., 2017)

Positional Encoding

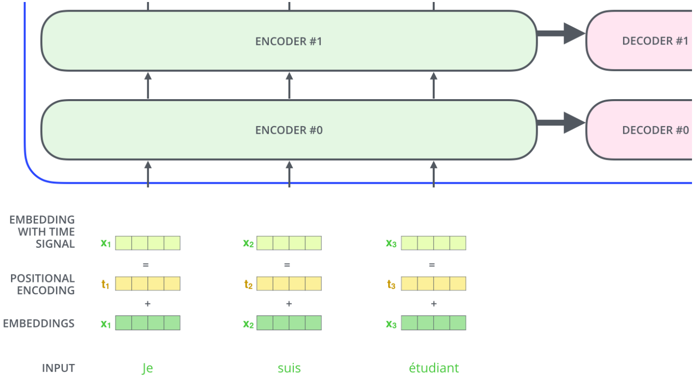


Figure: Attention heads (Alammar, 2018)

Positional Encoding

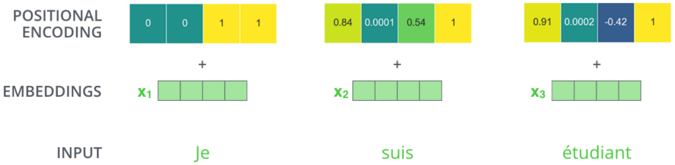


Figure: Adding positional encodings to embeddings (Alammar, 2018)

Add and Normalize

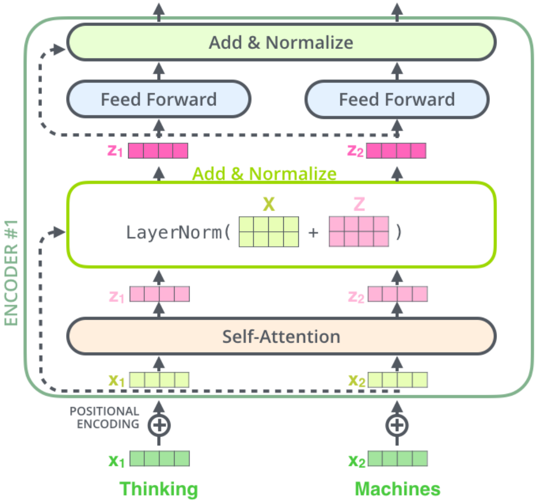


Figure: Add and Normalize (Alammar, 2018)

UPPSALA UNIVERSITET

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

Transformer

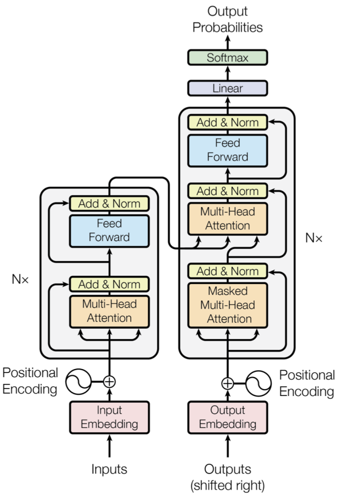


Figure: The Transformer Architecture (Vaswani et al., 2017)

UPPSALA UNIVERSITET

BERT

- Bidirectional Encoder Representations from Transformers
- Introduced in 2018/2019 in Devlin et al. (2017)

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

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

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
- Many new versions has come out RoBERTa, DistillBERT, AlBERT etc.

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
- Many new versions has come out RoBERTa, DistillBERT, AlBERT etc.
- Pre-trained on a large number of books

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
- Many new versions has come out RoBERTa, DistillBERT, AlBERT etc.
- Pre-trained on a large number of books
- Available both in English and Swedish (The National Library)

UPPSALA UNIVERSITET

BERT and transfer learning

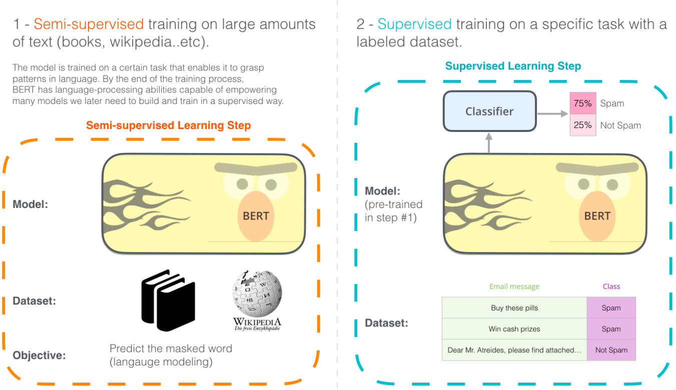


Figure: Using BERT for Transfer Learning (Alammar, 2019)

The BERT model

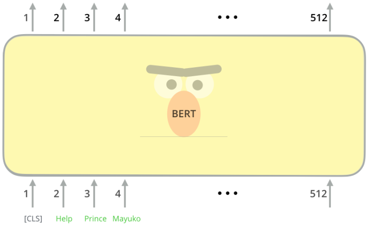


Figure: The BERT model (Alammar, 2019)

BERT Architecture

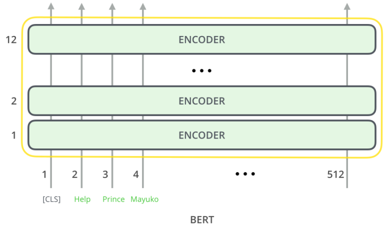


Figure: Opening up BERT (Vaswani et al., 2017)

Task 1: Masked Language Model

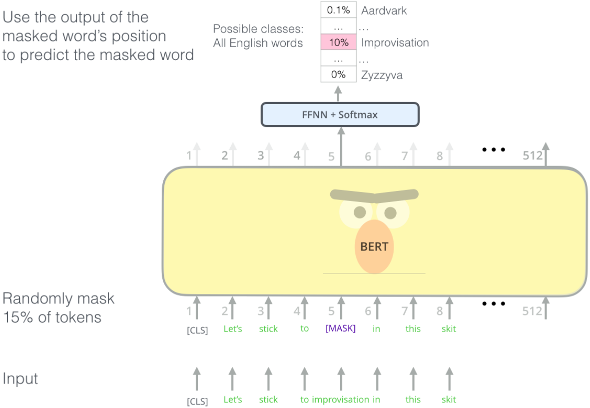


Figure: Masked Language Modeling (Alammar, 2019)

Next Sentence Prediction

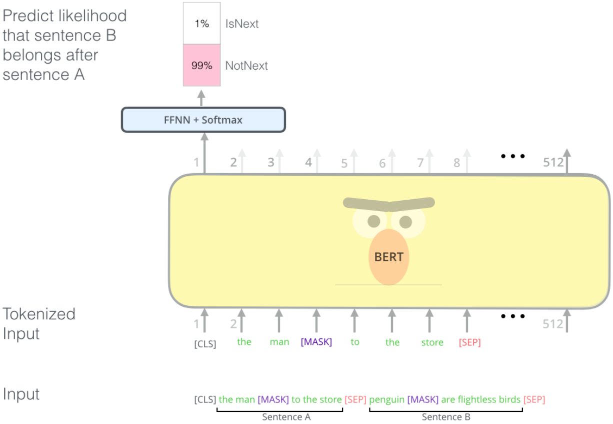


Figure: Next Sentence Prediction (Alammar, 2019)

Using BERT for Classification

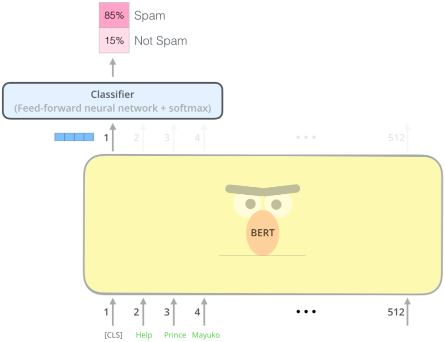


Figure: Using BERT for classification (Alammar, 2019)

BERT and Contextualized embeddings

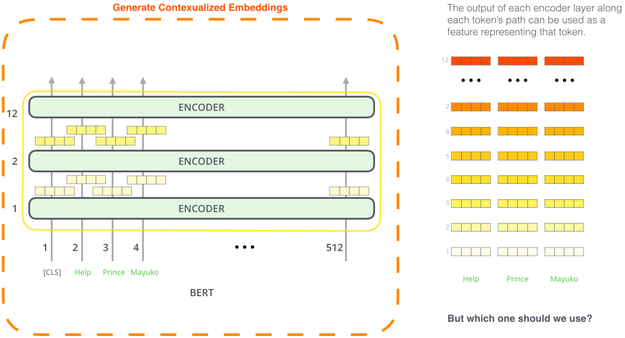


Figure: Contextualized Embeddings (Alammar, 2019)

Using Contextualized Embeddings

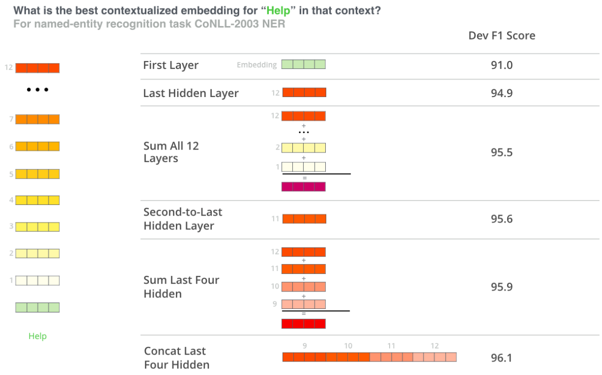


Figure: Using Contextualized Embeddings (Alammar, 2019)