

Using the Transformer

BERT (Devlin et al., 2018)



Learning goals

- Understand the use of the transformer encoder in this model
- Understand the architectural components

PREDECESSORS OF BERT

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.



2013

PREDECESSORS OF BERT

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework.

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

2013

01/2018

January 2018 - ULMFiT

The first transfer learning architecture (**Universal Language Model Fine-Tuning**) was proposed by **Howard and Ruder (2018)**.

An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A **Unidirectional contextual** model since no biLSTMs are used.

PREDECESSORS OF BERT

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

February 2018 - ELMo

Guys from **AllenNLP** developed a bidirectionally contextual framework by proposing ELMo (Embeddings from Language Models; **Peters et al., 2018**).

Embeddings from this architecture are the (weighted) combination of the intermediate-layer representations produced by the biLSTM layers.

2013

01/2018

02/2018

January 2018 - ULMFiT

The first transfer learning architecture (**Universal Language Model Fine-Tuning**) was proposed by **Howard and Ruder (2018)**.

An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A **Unidirectional contextual** model since no biLSTMs are used.

PREDECESSORS OF BERT

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the *word2vec* framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

February 2018 - ELMo

Guys from **AllenNLP** developed a bidirectionally contextual framework by proposing ELMo (Embeddings from Language Models; **Peters et al., 2018**).

Embeddings from this architecture are the (weighted) combination of the intermediate-layer representations produced by the biLSTM layers.

2013

01/2018

02/2018

06/2018

January 2018 - ULMFiT

The first transfer learning architecture (Universal Language Model Fine-Tuning) was proposed by **Howard and Ruder (2018)**.

An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A **Unidirectional contextual** model since no biLSTMs are used.

June 2018 - OpenAI GPT

Radford et al., 2018 abandon the use of LSTMs. They combine multiple Transformer decoder block with a standard language modelling objective for pre-training.

Compared to ELMo it is just **unidirectionally contextual**, since it uses only the decoder side of the Transformer. On the other hand it is **end-to-end trainable** (cf. ULMFiT) and embeddings do not have to be extracted like in the case of ELMo.

PREDECESSORS OF BERT

2013 - word2vec

Tomas Mikolov et al. publish four papers on vector representations of words constituting the word2vec framework

This received very much attention as it revolutionized the way words were encoded for deep learning models in the field of NLP.

February 2018 - ELMo

Guys from AllenNLP developed a bidirectionally contextual framework by proposing ELMo (Embeddings from Language Models; Peters et al., 2018).

Embeddings from this architecture are the (weighted) combination of the intermediate-layer representations produced by the biLSTM layers.

October 2018 - BERT

BERT (Devlin et al., 2018) is a bidirectional contextual embedding model purely relying on Self-Attention by using multiple Transformer encoder blocks.

BERT (and its successors) rely on the Masked Language Modelling objective during pre-training on huge unlabelled corpora of text.

2013

01/2018

02/2018

06/2018

10/2018

January 2018 - ULMFiT

The first transfer learning architecture (Universal Language Model Fine-Tuning) was proposed by Howard and Ruder (2018).

An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

A Unidirectional contextual model since no biLSTMs are used.

June 2018 - OpenAI GPT

Radford et al., 2018 abandon the use of LSTMs. They combine multiple Transformer decoder block with a standard language modelling objective for pre-training.

Compared to ELMo it is just unidirectionally contextual, since it uses only the decoder side of the Transformer. On the other hand it is end-to-end trainable (cf. ULMFiT) and embeddings do not have to be extracted like in the case of ELMo.