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.

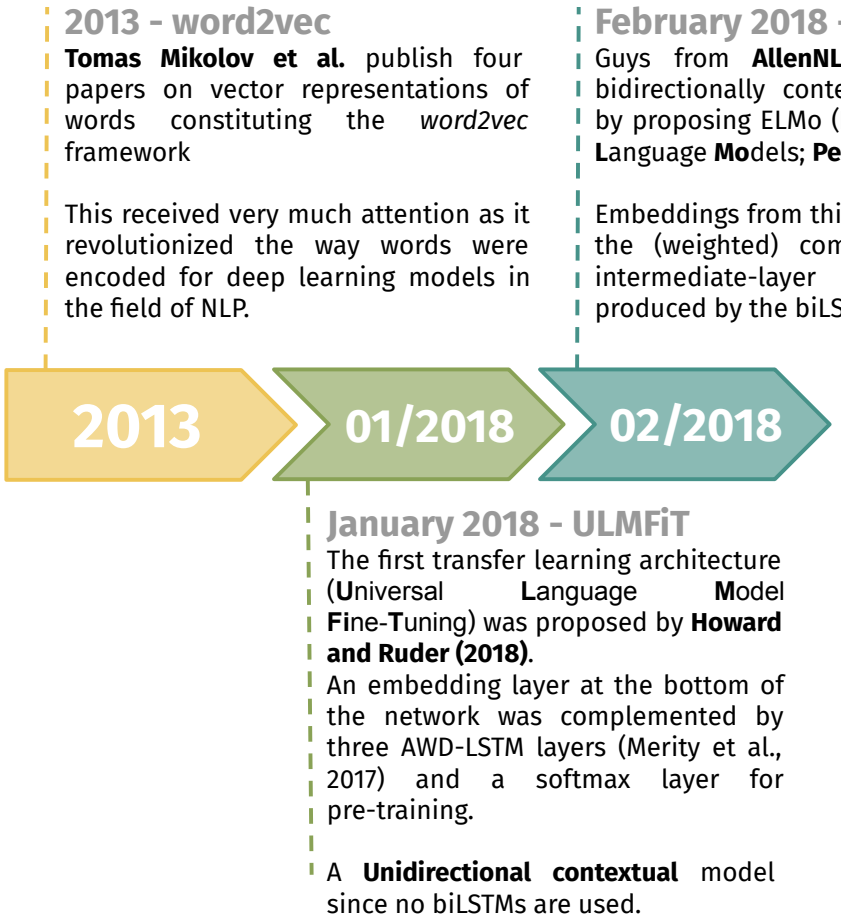
2013 → 01/2018

January 2018 - ULMFiT

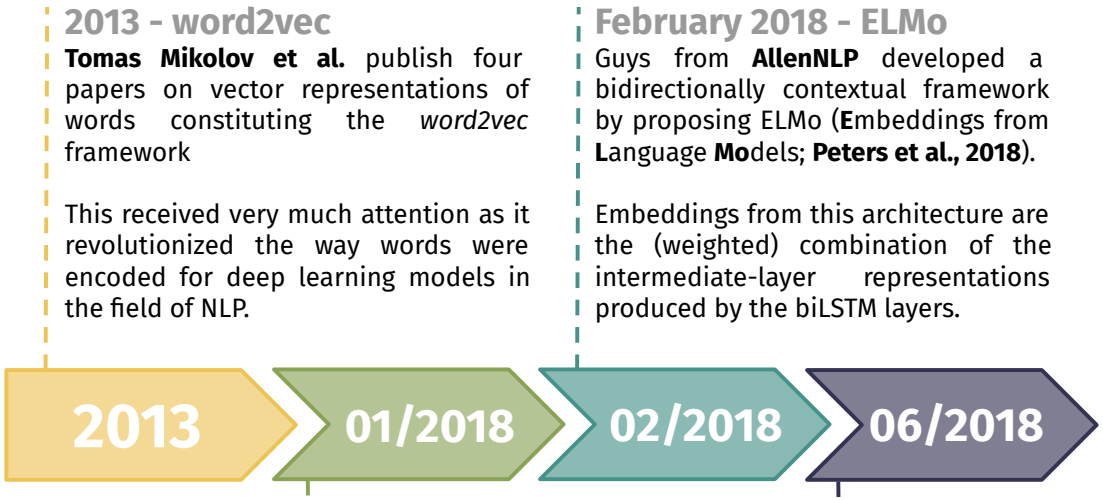
The first transfer learning architecture (**U**niversal **L**anguage **M**odel **Fi**ne-**T**uning) was proposed by **Howard and Ruder (2018)**.
An embedding layer at the bottom of the network was complemented by three AWD-LSTM layers (Merity et al., 2017) and a softmax layer for pre-training.

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

PREDECESSORS OF BERT



PREDECESSORS OF BERT



PREDECESSORS OF BERT