University of BRISTOL

Contextual Word Representations: A Contextual Introduction, Noah Smith 2020.

Chapter 11, Speech and Language Processing, Jurafsky & Martin (2021)

9.2 Contextualised Embeddings

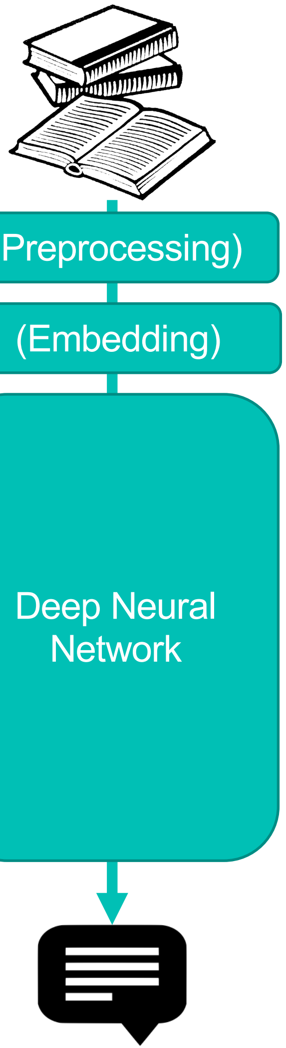
Edwin Simpson

Department of Computer Science,

University of Bristol, UK.

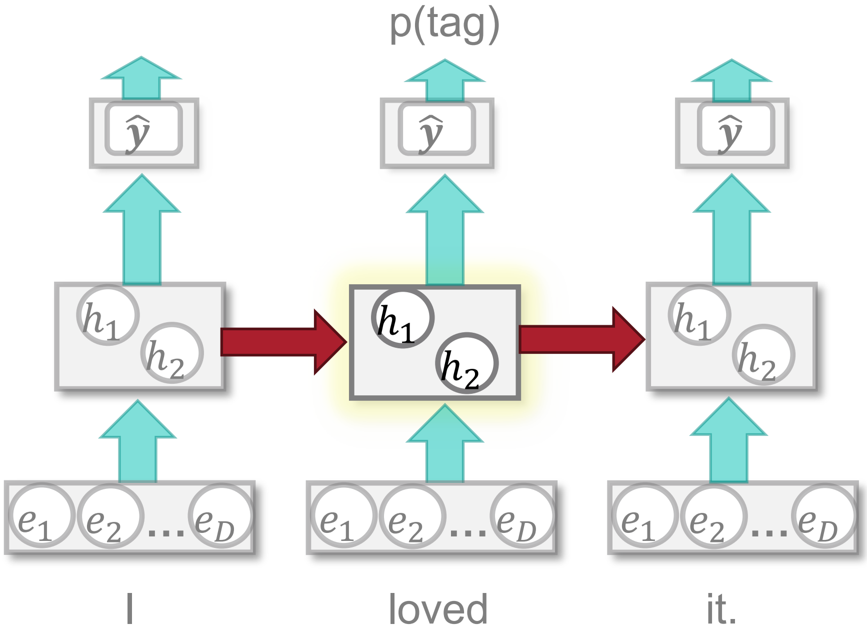
Hidden State Representations

- Deep neural networks learn **learn features** at various levels of abstraction.
- What are these features?
- The values of the hidden layers are vector representations of the text.



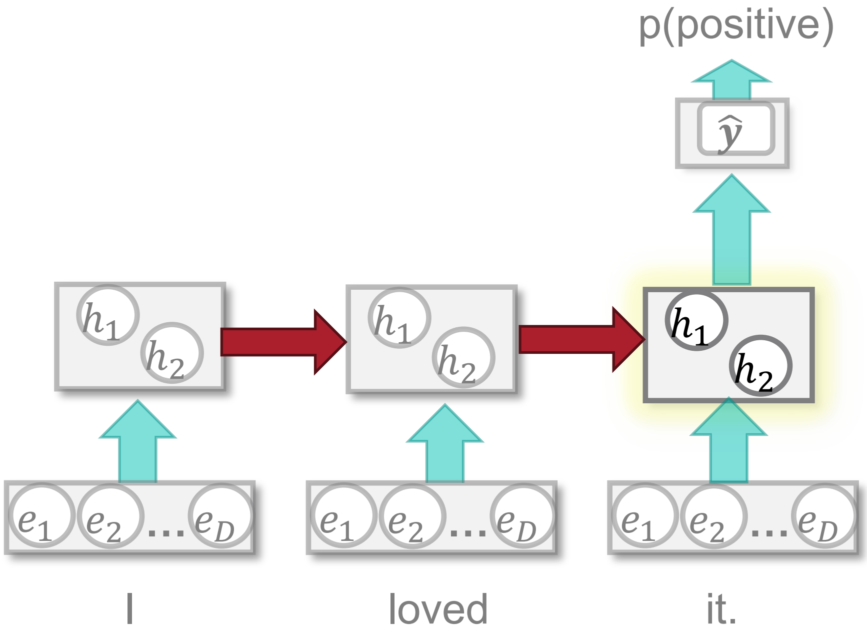
Hidden State Representations

- Deep neural networks learn **learn features** at various levels of abstraction.
- Examples of hidden layer representations:
- Vector representations of words in an RNN.
- The values of the hidden units in the RNN layer after processing the input 'loved'.



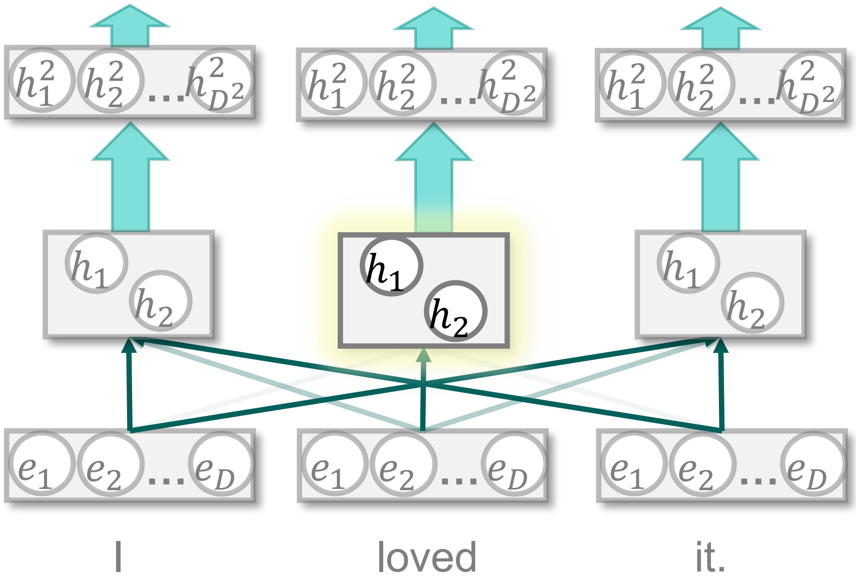
Hidden State Representations

- Deep neural networks learn **learn features** at various levels of abstraction.
- Examples of hidden layer representations:
- Vector representations of a sentence by an RNN.
- The values of the RNN hidden after processing the last token in the sequence →



Hidden State Representations

- Deep neural networks learn **learn features** at various levels of abstraction.
- Examples of hidden layer representations:
- Vector representations of words produced by a self-attention layer inside a transformer block
- Processing the input 'loved' →



Differences to Skip-gram Embeddings

- Can the hidden layer representations be considered as word or sentence embeddings?
- Hidden layers represent each token **in a specific context**, whereas skip-gram embeddings are fixed for each word type in a vocabulary.
- This is useful as many words don't have a constant meaning:
  - *... a **mouse** controlling a computer system in 1968.*
  - *.... a quiet animal like a **mouse.***
  - *… a small building in the **back.***
  - *A clear majority of senators **back** the bill.*

Section 18.1, Speech and Language Processing, 3rd edition draft, Jurafsky & Martin (2021).

Differences to Skip-gram Embeddings

- Skip-gram is trained to predict the context of a given word type…
- Unlike the LSTM or self-attention layer, the context window has a fixed size and word order is ignored…
- As a by-product of this task, skip-gram learns embeddings that are useful for many downstream tasks.
- So far, we've seen neural networks trained to do a downstream task like named entity recognition or sentiment analysis.
- Are there pretraining tasks for LSTMs or transformers that learn contextualised representations that are useful for downstream tasks?

Pretraining a Deep Neural Network

- **Language modelling**: given the first part of a text sequence, predict the next word.
- This is a **self-supervised learning** task, like skip-gram:

What do you think we mean by 'self-supervised' (compare with 'supervised' and 'unsupervised')?

[Semi-supervised Sequence Learning](). Dai and Le, 2015.
[Deep Contextualized Word Representations](), Peters et al. (2018).

Pretraining a Deep Neural Network

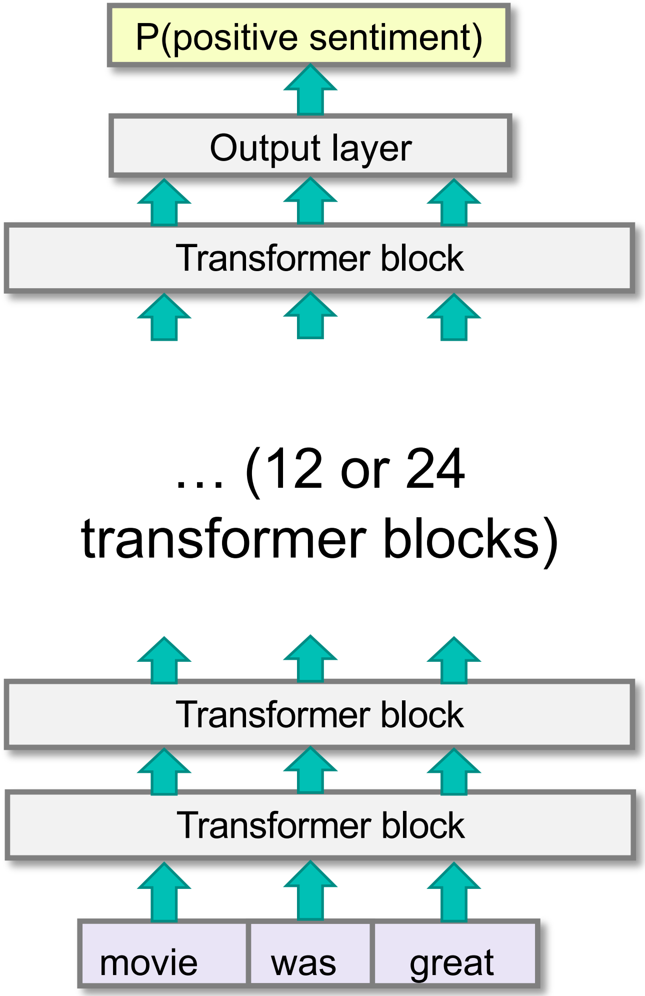
- **Language modelling**: given the first part of a text sequence, predict the next word.
- This is a **self-supervised learning** task, like skip-gram:
  - It does not require the data to be annotated to provide training labels.
  - It uses next word in an unlabelled text sequence as the label.
  - This means loads of data can be used to train the model.
- Problem: the context can only contain words that come **before** the current word, otherwise the network would 'see' tokens we need to predict later!

Semi-supervised Sequence Learning. Dai and Le, 2015.
Deep Contextualized Word Representations, Peters et al. (2018).

Google's BERT Model

- A large transformer model introduced in 2018
- Considers the whole sequence as context when encoding a word.
- Hence 'bidirectional' as context comes from before and after

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. Devlin et al., 2018.



Pretraining BERT

- **Masked language modelling:** mask out 15% of the words, then predict the masked words.

The man went to the [MASK] to buy a [MASK] of milk

shop

bottle

Which words are missing?
What did you need to know to guess them?

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Devlin et al., 2018.

Pretraining BERT

- **Next Sentence Prediction:** does the second sentence follow the first?

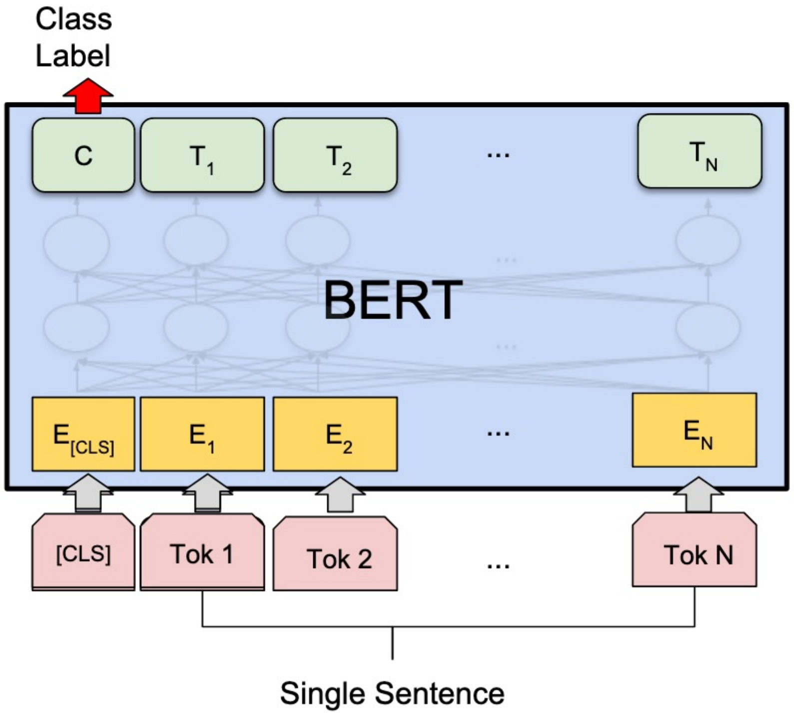
[CLS] The man went to the shop. [SEP] He bought some milk.

True

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. Devlin et al., 2018.

BERT for Text Classification

- Special [CLS] token
- Embedding for [CLS] learns to represent the whole sentence
- Final [CLS] embedding used as input to a classifier

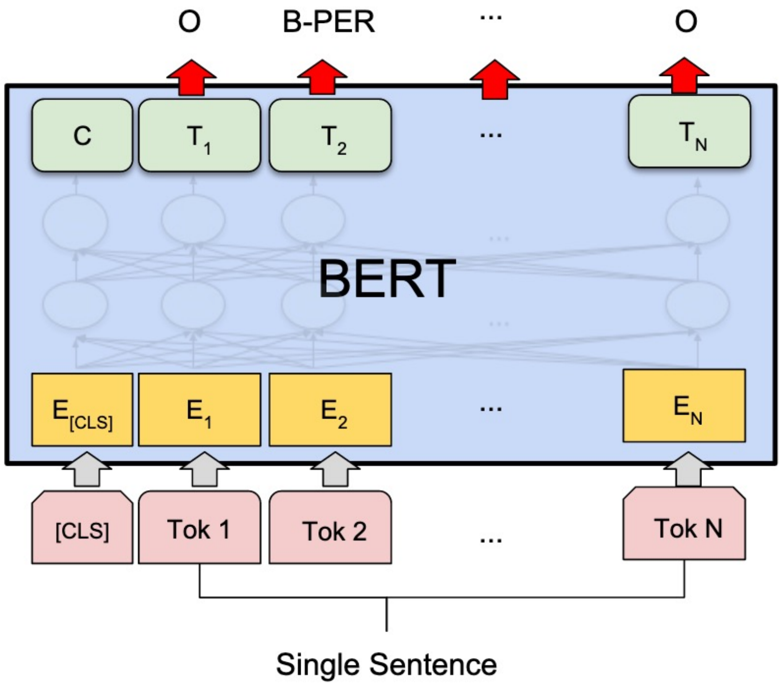


[CLS] The man went to the shop to buy a bottle of milk

Figure from BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. Devlin et al., 2018.

BERT for Sequence Tagging

- Outputs of the last transformer blocks are contextualised word embeddings
- 768-dimensions



[CLS] The queen … blue.

Figure from BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.. Devlin et al., 2018

BERT is a **Large** Language Model

- English data: Wikipedia (2.5B words) + publicly-available books (800M words) and 30,000 vocabulary.
- Google has trained BERT on >70 languages.
- Training time: 1M steps (~40 epochs), four days on Google’s TPU.
- Architecture of the ‘BERT-base’ model: 12 transformer blocks, each containing 12 self-attention heads with feedforward layers
- 110 million parameters in total.

BERT Performance

- Error reductions compared to non-contextualised embeddings:
  - Answering questions about content in a piece of text – 50% (SQuAD).
  - Using common sense to infer what happens next – 67% (SWAG).
- BERT reduces errors dramatically on complex text understanding tasks.
- Trade-off is that computational/memory costs are much higher, not always worth it.

BERT Variants

- Plethora of newer pretrained transformer models, including:
- Multilingual BERT: 104 languages trained with Wikipedia.
- Distilbert: 40% fewer parameters but retains 97% performance.
- RoBERTa: better performance from an enlarged dataset and optimised pretraining procedure.
- XLM-RoBERTa: cross-lingual version of RoBERTa for >100 languages.

Summary

- The hidden states of a transformer or LSTM can be used as contextualised word embeddings.
- These vectors capture context and disambiguate words.
- Transformers are pretrained on (masked) language modelling with huge amounts of unlabelled text.
- The pretrained models can then map text to contextualised embeddings for use in downstream tasks.