BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

TAYLOR DOWNEY

COEN 296: NLP

11/03/20

What is BERT?

- A language model that stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers
- Created and published by Google AI Language Research Lab in 2018
- Achieved state of the art performance on several NLU datasets including
  - SQuAD – Stanford Question Answering Dataset
  - GLUE – General Language Understanding Evaluation
  - SWAG – Situations with Adversarial Generations
- Won the Best Long Paper Award at the 2019 NAACL (North American Chapter of the Association for Computational Linguistics)

The sequence-to-sequence architecture

- A neural net that transforms a given sequence of elements into another sequence
  - Typical application is Machine Translation -> convert one language to another
- Encoder takes the input sequence and maps it into a higher dimensional space (n-dimensional vector)
- Decoder takes the abstract vector and turns it into an output sequence
- RNNs typically used are LSTMs or GRUs

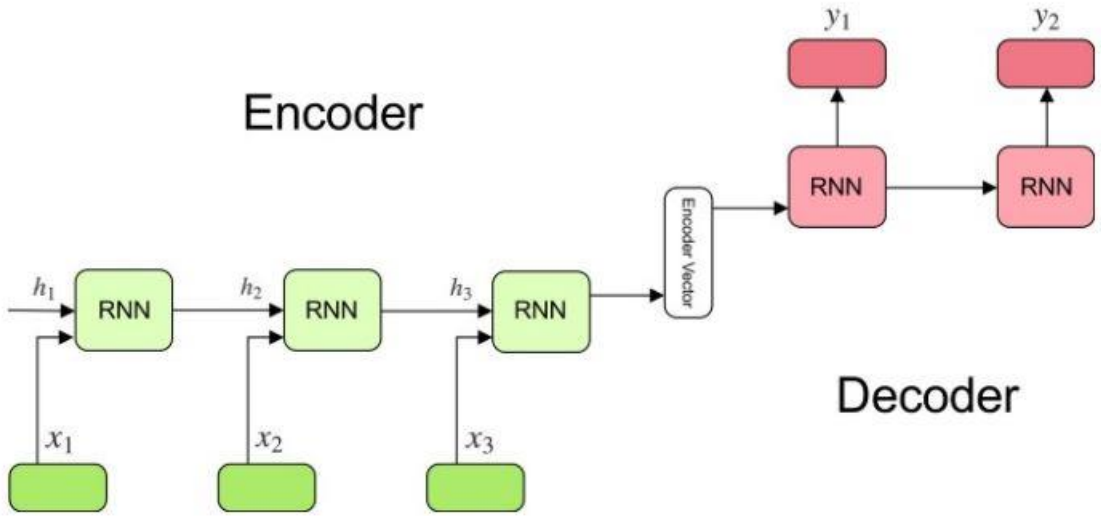


Figure from [5]

Attention Mechanism

- Architecture of any RNN suffers from the vanishing gradient problem
  - During backpropagation, a network's weights receive an update proportional to the partial derivative of the error function with respect to the current weight in each iteration of training. With deep networks, the gradient can become so small that the weights change very slowly or not at all

- As a result, the RNN can have a difficult time with long sentences because it ‘forgets’ the previous input.
  - The coronavirus pandemic has spread across 175 countries and **it** is beginning to spread rapidly again in Europe.

- Words in a sentence can be related to each other even if they are not next to each other. Same goes for sentence to sentence or even paragraph to paragraph

- To give the networks better memory, allow each input word to contribute to the output, rather than the strictly sequential process done previously

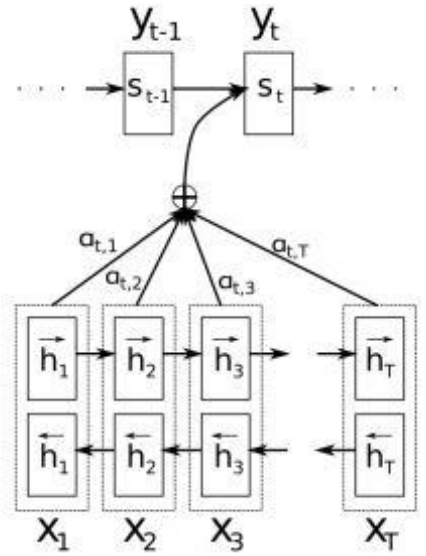


Figure from [6]

What is a Transformer?

- Google Research published a paper in 2017 called ‘Attention Is All You Need’ - Deep learning model built off the ideas of ‘attention’ but without using an RNN

- As a result, the Transformer does not require sequential data to be processed in order and does a better job at handling long input sequences.

- Four main components:
  - Input Embedding
  - Positional Encoding
  - *Multi-Head Self Attention*
  - Feed Forward Network

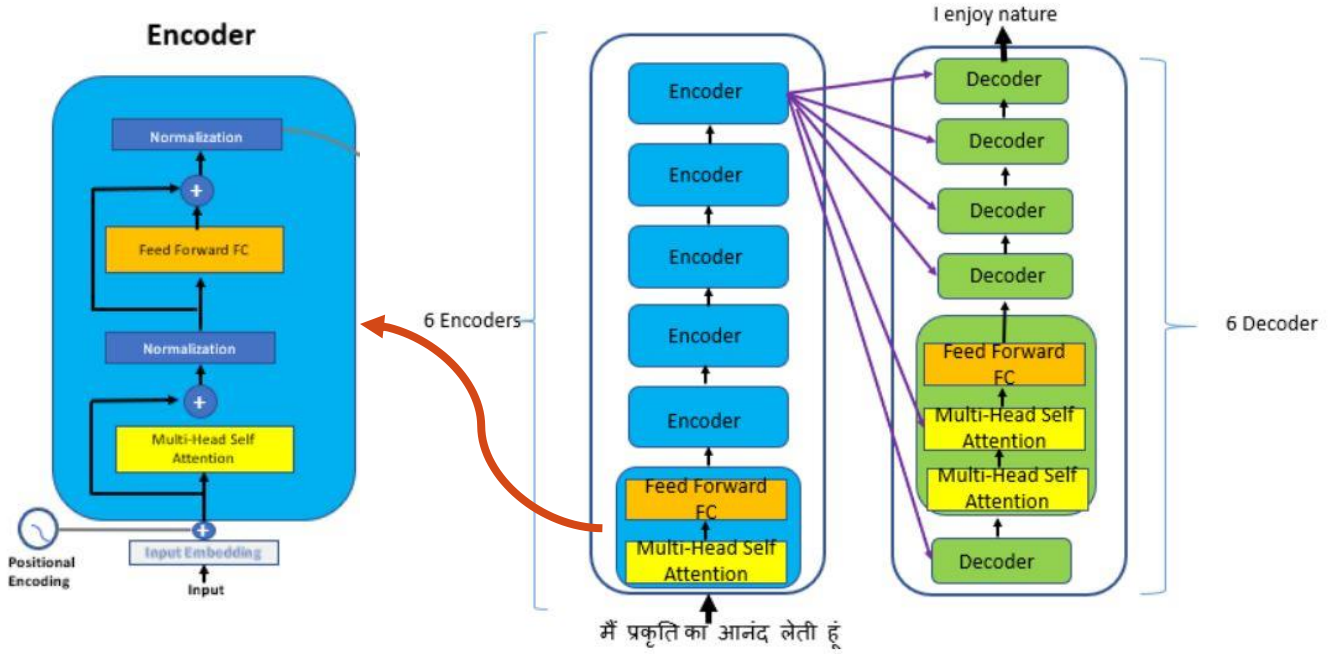


Figure from [4] High-level Transformer architecture

Multi-Head Self Attention

- Self attention is an attention mechanism that essentially learns how words correspond and interact with each other

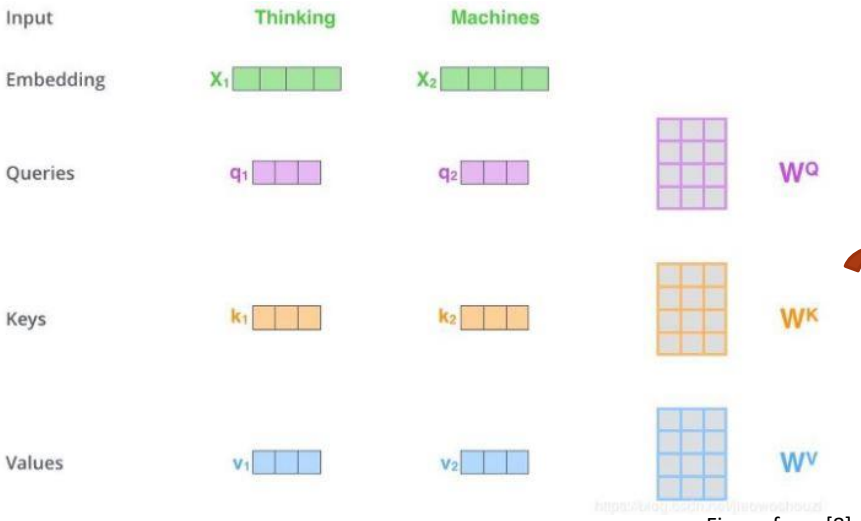


Figure from [8]

Transform word embeddings into 3 vectors: Queries, Keys, Values

Apply transformations to vectors in order to extract which words are most important and how they relate to each other

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Perform this operation 8 times in parallel. In theory, this creates a more robust understanding of the input

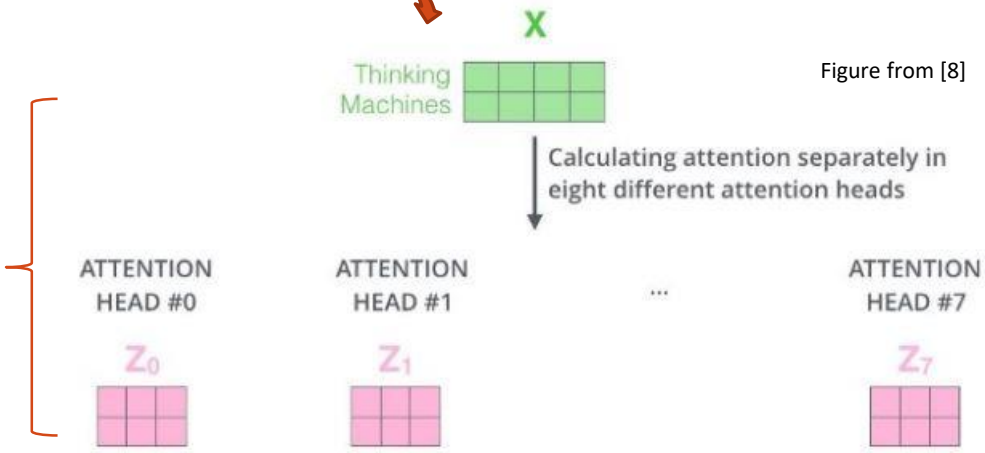


Figure from [8]

What does BERT do?

- BERT's key innovation was applying the bidirectional training of Transformer's to language modeling.
- Recall that a statistical language model is a probability over sequences of words. Given a sequence of words, assign a probability to the whole sequence

  - n-gram model
  - Bidirectional model
  - Maximum Entropy model

- Since BERT is a language model, it only uses the Transformer encoder.
- There are two training phases: Pre-training and Fine-Tuning

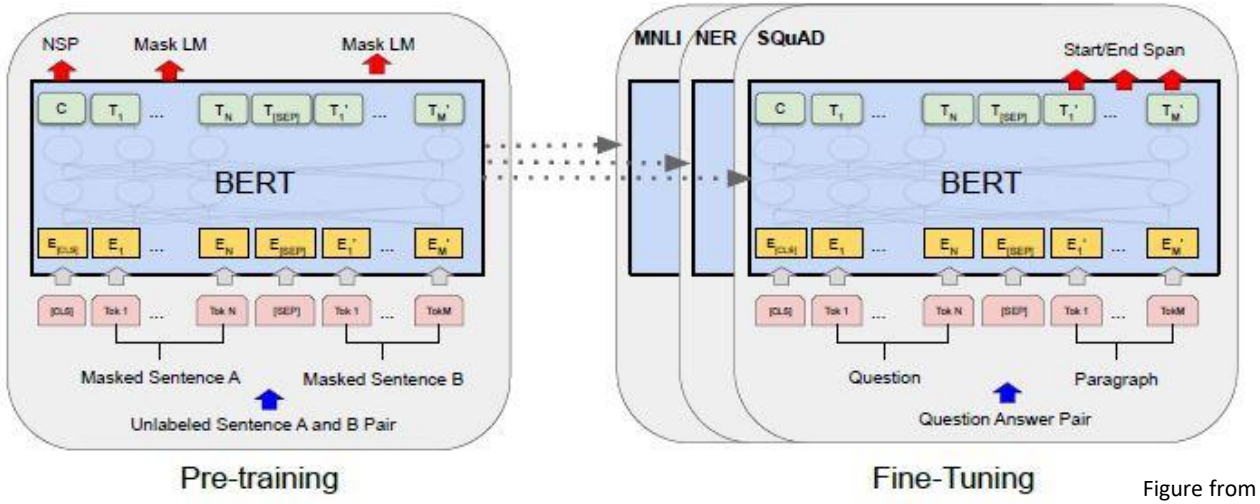


Figure from [1]

Pre-Training BERT

- The pre-training phases consists of two unsupervised predictive tasks -> learn from unlabeled data

Masked Language Model (MLM)

- Before feeding word sequences into model, 15% of the words in each sequence are replaced with a [MASK] token.
- Model then attempts to predict the original value of the masked words based on the context provided by the other non-masked words in the sequence
- Prior language models are trained left-to-right, right-to-left, or some shallow concatenation of the two.
- Using MLM, BERT becomes a deep bidirectional model and therefore much more powerful

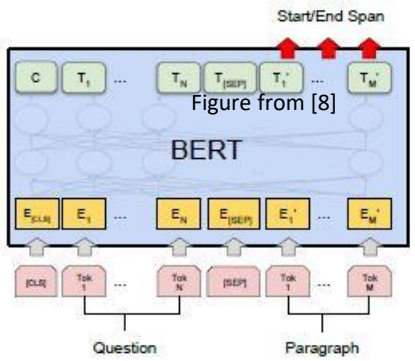
Next Sentence Prediction (NSP)

- Many tasks in NLU are based on understanding the relationship between two sentences, which is not directly captured by language modeling.
- The model receives pairs of sentences as input and learns to predict if the second sentence in the pair is the subsequent sentence in the original document
- During training, 50% of the inputs are pairs while the other 50% are randomly chosen

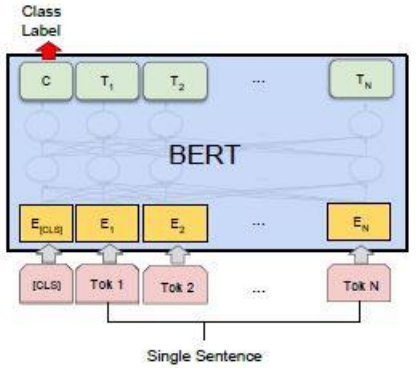
Combine to create loss function!
When training the BERT model, MLM and NSP are trained together with the goal of minimizing the combined loss function of the two strategies

Fine-Tuning BERT

- Once you have a trained language model, you can fine tune it for a specific task
- Task specific inputs and outputs are pushed into BERT and all hyper parameters are tuned end to end
- *Sentiment Analysis*: add a classification layer on top of transformer output
- *Question Answering*: Add two extra vectors that mark the beginning and ending of an answer
- *Named Entity Recognition*: add a classification layer that receives the output vector of each token



(c) Question Answering Tasks: SQuAD v1.1



(b) Single Sentence Classification Tasks: SST-2, CoLA

Figure from [8]

Both figures from [1]

BERT Takeaways

- Model size matters
  - BERT Base
    - 12 Encoders, 768 hidden units in feed forward network, 12 attention heads
  - BERT Large
    - 24 Encoders, 1024 hidden units in feed forward network, 16 attention heads

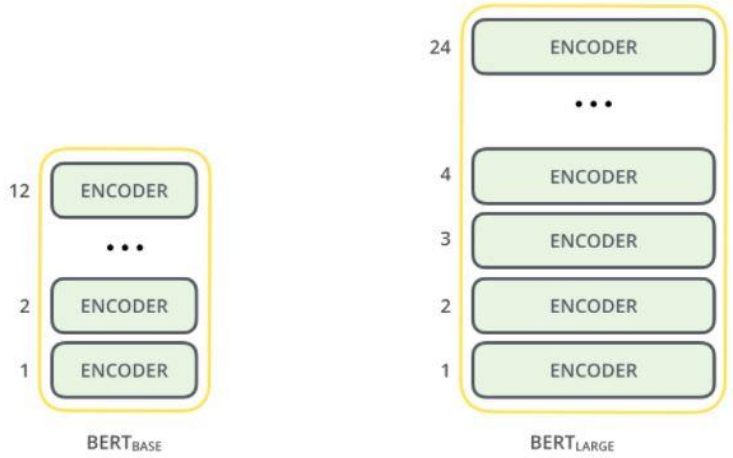


Figure from [10]

- With enough training data, more training *steps* yields higher accuracy
  - When base model was trained on 1M steps rather than 500k steps, accuracy improved by 1%

- Bidirectional training converges much slower than other approaches

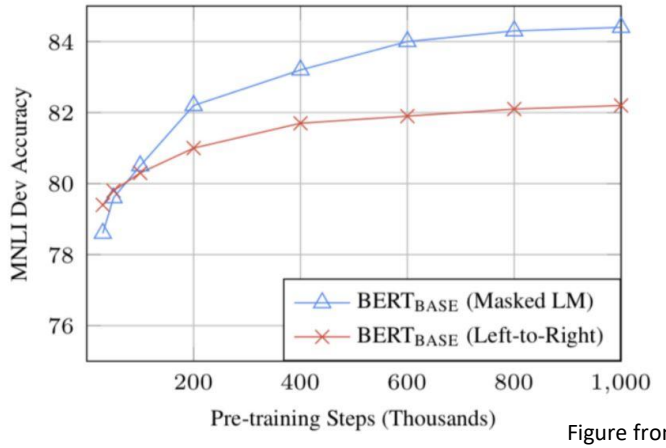


Figure from [1]

References

1. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018
2. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010
3. https://en.wikipedia.org/wiki/BERT_(language_model)
4. https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
5. https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346
6. http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/
7. https://en.wikipedia.org/wiki/Vanishing_gradient_problem
8. https://towardsdatascience.com/breaking-bert-down-430461f60efb
9. https://medium.com/inside-machine-learning/what-is-a-transformer-d07dd1fbec04
10. http://jalammar.github.io/illustrated-bert/
11. https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)