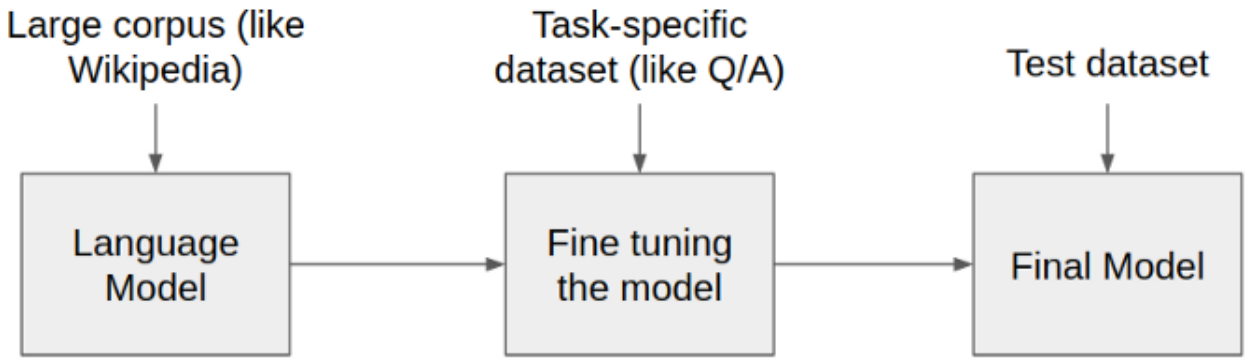
Pre-trained Language Models: Simplified

Sesame street of the NLP world

Prakhar Ganesh Dec 17, 2019 · 4 min read ★

What are pre-trained language models?

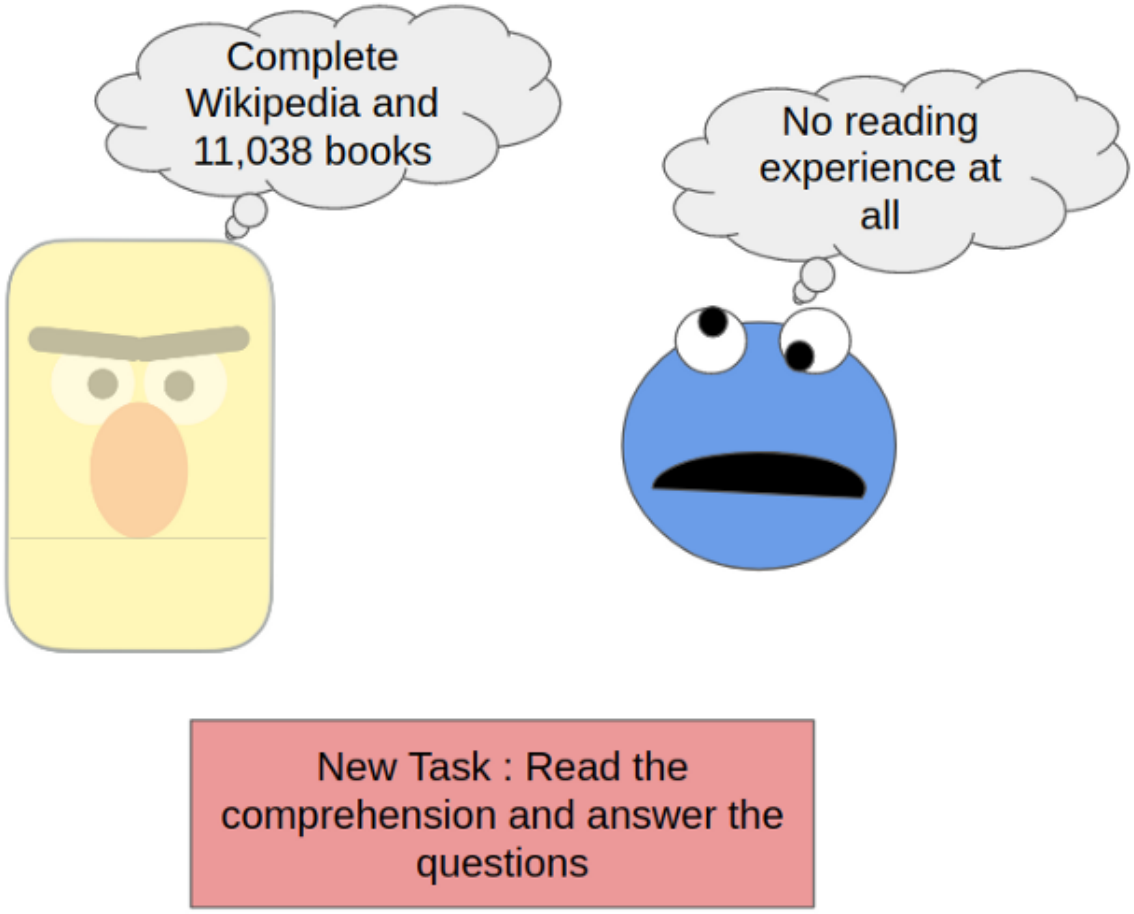
The intuition behind pre-trained language models is to create a black box which **understands** the language and can then be asked to do any specific task in that language. The idea is to create the machine equivalent of a 'well-read' human being.



The language model is first fed a large amount of unannotated data (for example, the complete Wikipedia dump). This lets the model learn the usage of various words and how the language is written in general. The model is now transferred to an NLP task where it is fed another smaller task-specific dataset, which is used to fine tune and create the final model capable of performing the aforementioned task.

Why are they better than task-focused models?

In one sentence : *they are better read!!* A model which trains only on the task-specific dataset needs to both understand the language and the task using a comparatively smaller dataset. The language model on the other hand already understands the language since it has 'read' large language dumps during pre-training. Thus the language model can directly fine tune itself to match the required task and performs better than the existing SOTA.



Embedding vs Fine tuning

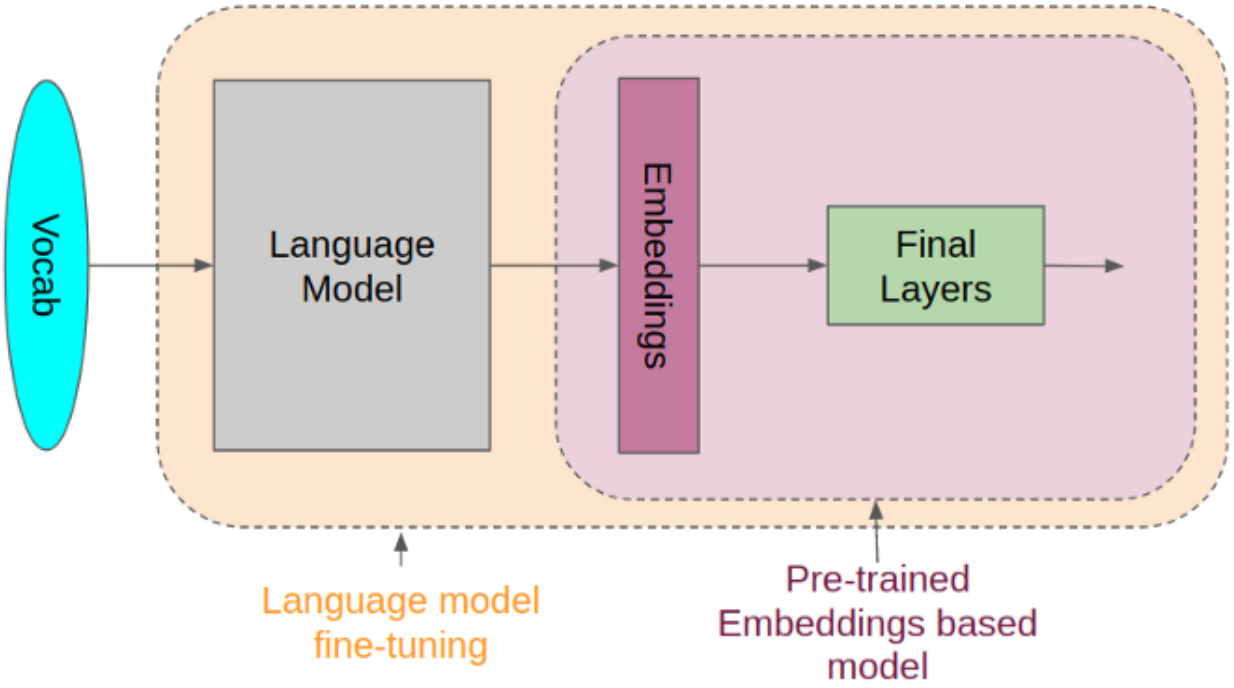
Every word in NLP needs to be represented mathematically in order to let machines do further processing. You can gain more intuitions on this through my earlier blog on Distributed Vector representation…

Distributed Vector Representation : Simplified

Arguably the most essential feature representation method in Machine Learning

towardsdatascience.com

Many different algorithms have been proposed to create these embeddings, by pre-training the models on a separate larger dataset to capture the essence of the language. For example, Word2Vec embeddings gained incredible popularity and these embeddings were directly used for a number of tasks across NLP.

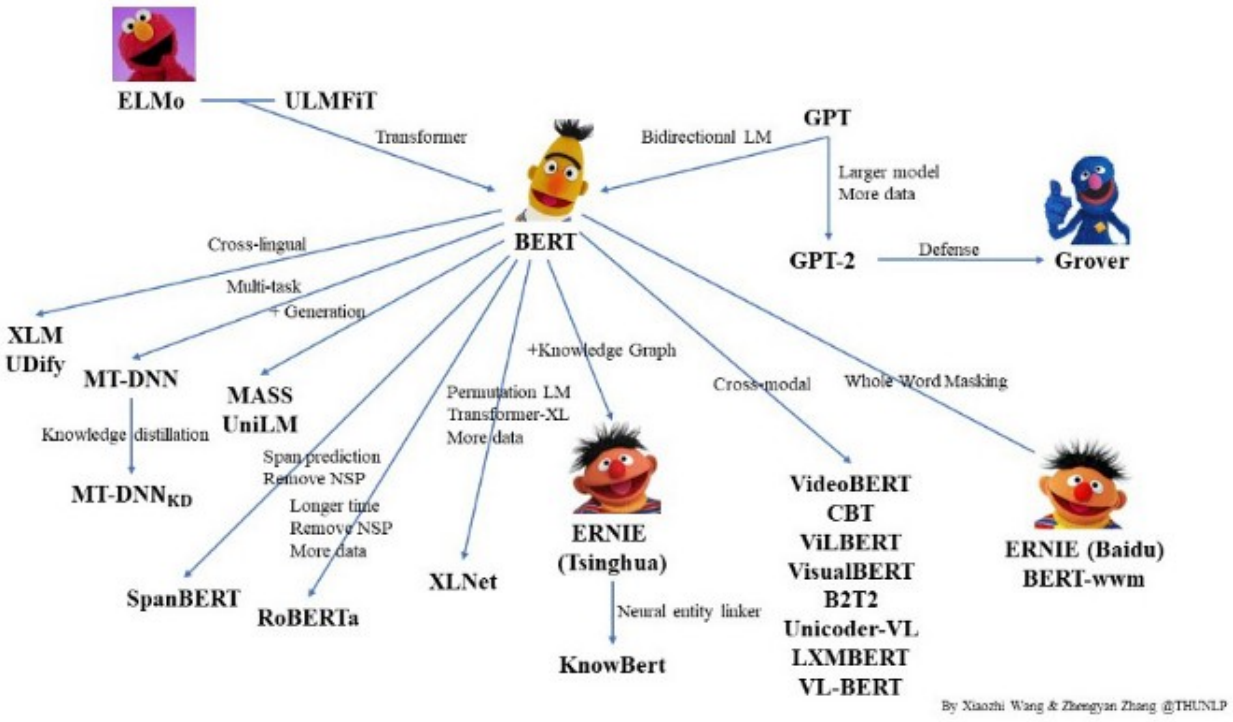


However, these word representations are learned in a generalized context and do not represent task-specific information. This is where the fine-tuning part of a language model comes into account. Directly using the pre-trained embeddings can decrease the overall model size but forces us to only utilize generalized word representations. Language model fine-tuning on the other hand allows the user to fine-tune these word embeddings/representations by training on the task-specific dataset.

For example, the representation of the word ‘current’ might have good relation with both ‘news’ and ‘electricity’ in a generalised context. However, for a specific task which talks about electric circuits, allowing the model to fine-tune the word representations such that ‘current’ and ‘electricity’ match better can help improve the model performance.

BERT

BERT (Bi-directional Encoder Representations from Transformers) was proposed by Google last year and was able to *single-handedly achieve SOTA performances on 11 separate NLP tasks!!* It has since been the source of multiple language models spawning from it.



The model owes it's success mainly to the training methods proposed in the paper. The two training protocols, namely 'Masked LM' and 'NSP : Next Sentence Prediction' (which was later improved to 'SOP : Sentence Order Prediction'), helps BERT learn from the huge language corpus available to it.

The 'Masked LM' task is implemented by masking 15% of the word randomly in every sentence and training the model to predict them. The 'SOP' task is a classification task with two sentences input and the model is expected to recognize the original order between these 2 sentences, which increases it's document level understanding. The impact these training tasks create and the internal working of BERT requires a more detailed analysis which I will not go into right now.

What's next?

While different variations of language models are starting to dominate across various NLP tasks, everybody is starting to realize two important facts. One, although the original BERT model had the right idea, it was poorly trained and thus has incredible untapped potential. And two, the monstrous size of these pre-trained language models like BERT are a big hurdle towards future research and deployability of these models. It seems like these two major questions need to be answered before this field can move forward.

This blog is a part of an effort to create simplified introductions to the field of Machine

Learning. Follow the complete series here

Machine Learning : Simplified

Know it before you dive in

towardsdatascience.com

Or simply read the next blog in the series

Types of Convolution Kernels : Simplified

An intuitive introduction to different variations of the glamorous CNN layer

towardsdatascience.com

References

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[2] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[3] Peters, Matthew E., et al. "Deep contextualized word representations." arXiv preprint arXiv:1802.05365 (2018).

[4] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

[5] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

Machine Learning Naturallanguageprocessing Language Model Surveys Data Science