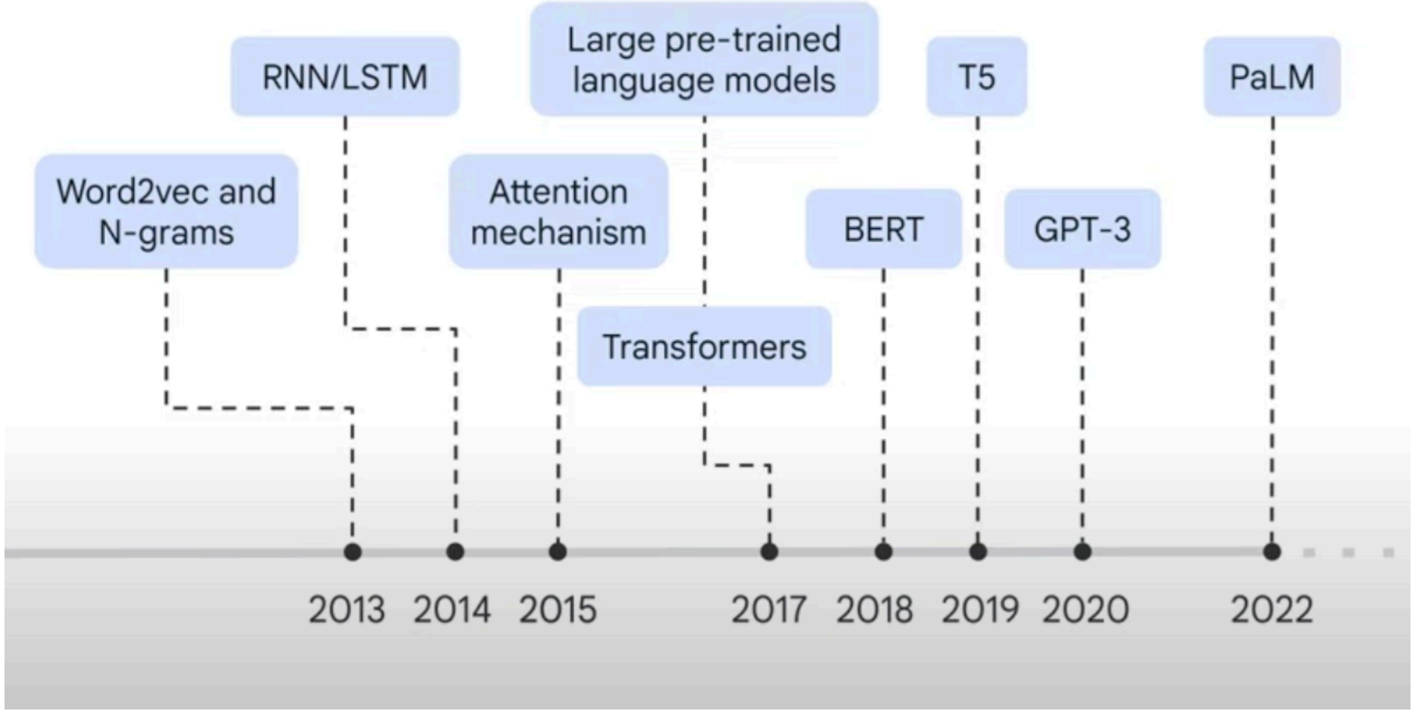
Natural Language Processing

Dr. Manuel Pita & Zuil Pirola

UNIVERSIDADE
LUSÓFONA

Timeline



https://www.youtube.com/watch?v=t45S_MwAcOw&t=578s

Evolution of NLP Models

Before BERT:

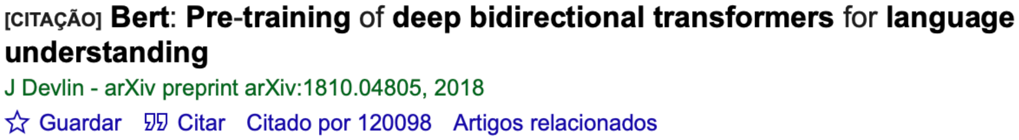
- Traditional NLP models: Word2Vec, GloVe (**static** word embeddings).
- **Unidirectional** models: LSTM, GRU, and older transformer models.

BERT's Innovation:

- **Bidirectional** training (contextual understanding from both directions).
- Transformer architecture for parallel processing.

Understanding BERT

- BERT stands for **Bidirectional Encoder** Representations from **Transformers**.
- Introduced by Google AI in 2018.



[CITAÇÃO] **Bert**: **Pre-training** of **deep bidirectional transformers** for **language understanding**

J Devlin - arXiv preprint arXiv:1810.04805, 2018

☆ Guardar ❞ Citar Citado por 120098 Artigos relacionados

- Used for tasks like translation, sentiment analysis, question answering, and more.
- Key innovation: bidirectional context, enabling better semantic understanding.

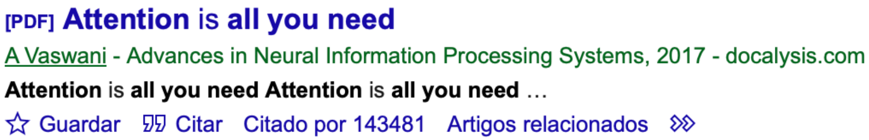
Understanding BERT

Let’s break this down into three key components:

- Transformers: The backbone architecture.
- Encoder: The part of the transformer that BERT uses.
- Bidirectional: BERT’s unique approach to context understanding.

Transformers

- Introduced by Vaswani et al. in 2017.



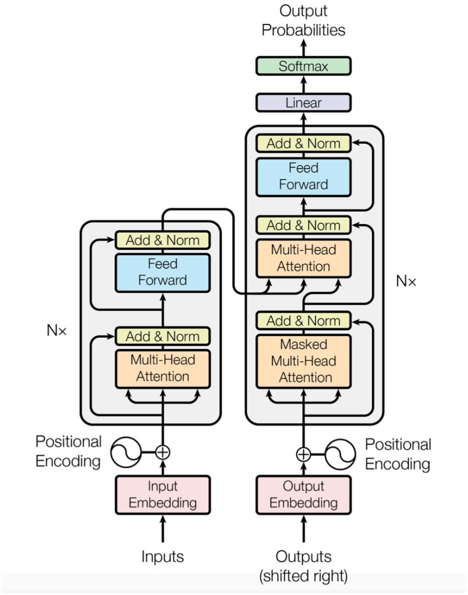
[PDF] **Attention** is **all you need**

A Vaswani - Advances in Neural Information Processing Systems, 2017 - docalysis.com

Attention is **all you need Attention** is **all you need** …

☆ Guardar ❞ Citar Citado por 143481 Artigos relacionados ≫

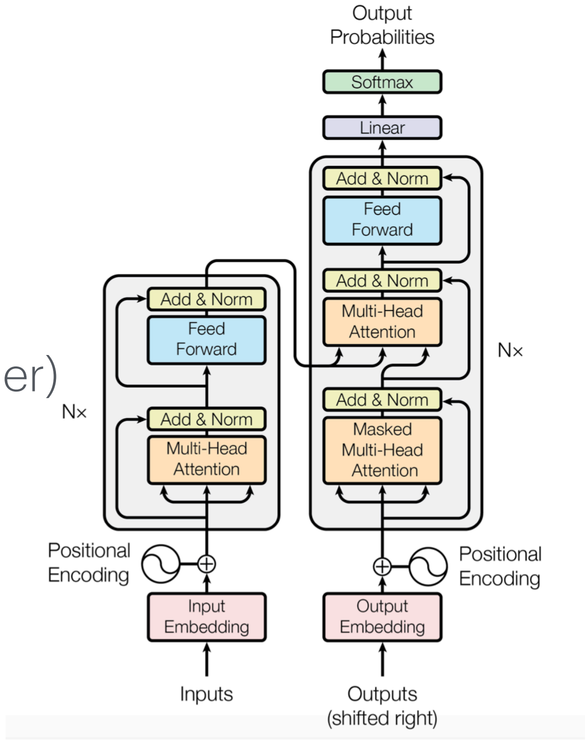
- Revolutionized NLP with self-attention mechanisms.



Encoder

Steps in the Encoder:

- Input tokens are converted into **embeddings**.
- **Positional Encoding** is added to retain word order information. (Order)
- Tokens pass through multiple **self-attention** and FFNN layers



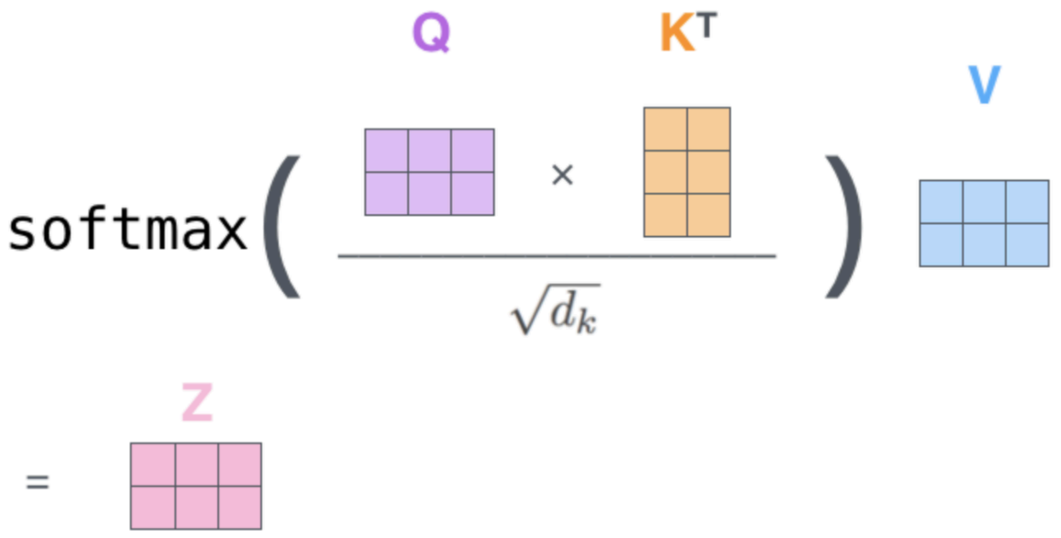
Encoder

Self-attention:

- mechanism allows it to capture **contextual relationships**.
- **Focus** in specific part!

    Ex: The black cat ate the mouse. (Translate)

Encoder



The self-attention calculation in matrix form

https://www.linkedin.com/pulse/self-attention-why-we-divide-score-square-root-key-vector-yadav-/

Bidirectional

BERT processes text by considering the **entire context** (both left and right) simultaneously.

- Masked Language Modeling (MLM)
    - Randomly masks ~15% of tokens in the text.

Ex:

Input: "The cat is [MASK] on the roof."

Output: Predicts "sitting" based on the context from **both directions**.

Bidirectional

- **Better Context Understanding**:

  Captures meaning in ambiguous cases.

  Example: "bank" in:

  "He went to the bank to sit."

  "He went to the bank to withdraw money.”

- Bidirectional models predict tokens based on full context, which **doesn't align with generation tasks**.

Bidirectional

What is Next Sentence Prediction (NSP)?

- NSP is a pretraining task in BERT to help the model understand the **relationship between two sentences**.

  Determine if a given sentence **B** logically follows another sentence **A**.

  [CLS] A [SEP] B [SEP]

- Understand contextual relationships between sentences. (**QA**)

Results

Models with:

- Context token vectors

  Each token or word: **768 dimensions** (for BERT-base) or **1024 dimensions** (for BERT-large).

- Sentence vectors

  A sentence or paragraph: Size depends on the aggregation method (e.g., averaging, using the **[CLS] token**).

Pre-training Vs Fine-tuning

In Pre-training, model learns from **large**, **unlabeled** text datasets.

Tasks:

- Masked Language Modeling (MLM): Predict masked words.
- Next Sentence Prediction (NSP): Predict sentence pairs' relationship

https://huggingface.co/

Pre-training Vs Fine-tuning

In Fine-tuning, load pre-trained model and adapt.

- Adapting the pretrained model to specific tasks (e.g., sentiment analysis, QA).
- Requires labeled data for the task.
- Pretrained models available for transfer learning.

https://huggingface.co/

Advantages and Disadvantages of BERT

Advantages:

- **Contextual Word Representations:** Unlike static embeddings (e.g., Word2Vec), BERT understands words based on context, significantly improving task performance.
- **State-of-the-Art Performance:** Achieved top results in multiple NLP benchmarks (e.g., GLUE, SQuAD).
- **Transfer Learning:** Allows fine-tuning on specific tasks with smaller datasets, saving time and resources.
- **Versatility:** Can be applied to a wide range of NLP tasks like question answering, sentiment analysis, and named entity recognition.

Advantages and Disadvantages of BERT

Disadvantages:

- **High Computational Cost:** The large model size (with millions of parameters) makes BERT resource-intensive for both training and inference.
- **Slow Inference:** Fine-tuned BERT models are slow compared to simpler models, especially for real-time applications.
- **Limited by Pre-training Data:** If the pre-training data does not include specific domain knowledge, BERT may perform poorly on specialized tasks.

Obrigado!