TRANSFORMERS

Pedro Henrique Barbosa de Almeida
Thomas Palmeira Ferraz

Imagens de Hasbros Studios

INTRODUÇÃO

Recapitulação de "*word embedding*" e de modelos pré-*Transformers*.

01

TRANSFORMERS

Surgimento, motivação, "atenção" e performance.

02

BERT

Representação Bidirecional de Contexto

03

ESTADO-DA-ARTE

Um novo marco a cada mês

04

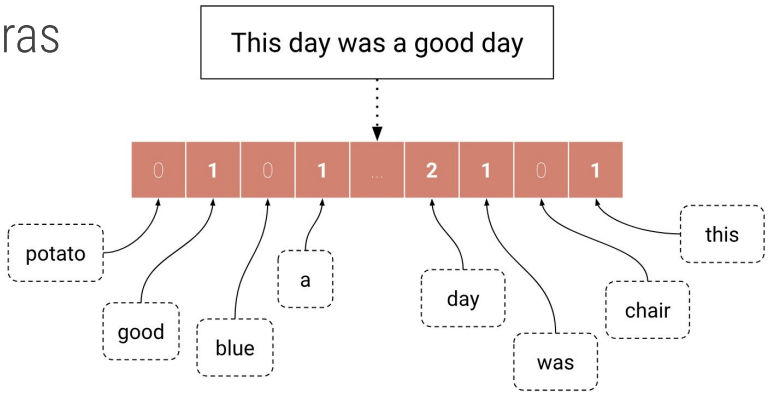
01

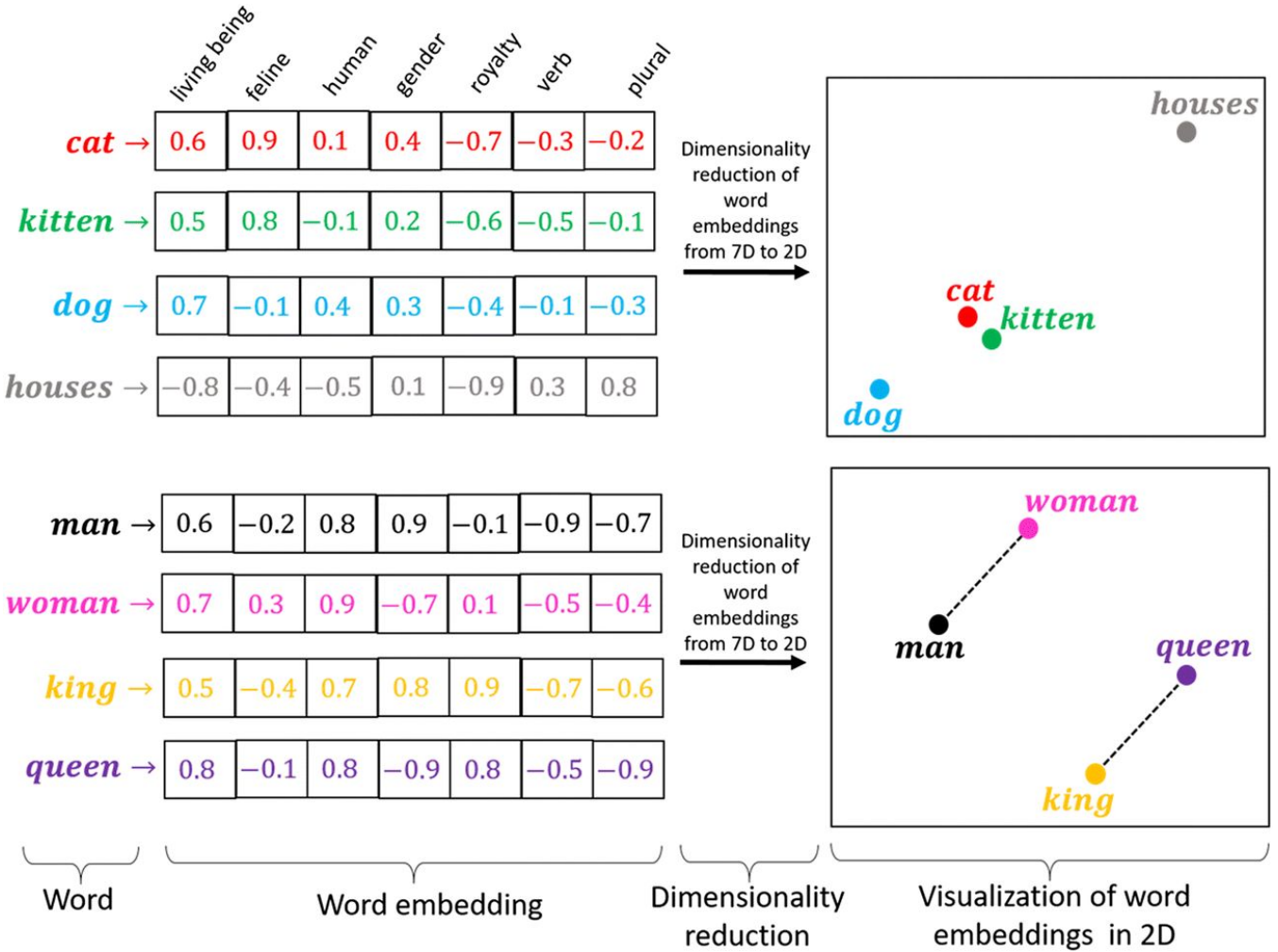
INTRODUÇÃO

Recapitulação de "*word embedding*" e de modelos pré-*Transformers*.

Introdução

- Problema fundamental do Processamento de Linguagem Natural: como transformar palavras em números?
  - *One-hot encoding / Bag-of-Words:*
    - Matrizes esparsas.
  - *Word Embeddings:*
    - Vetores densos;
    - Aprendizado de pesos.

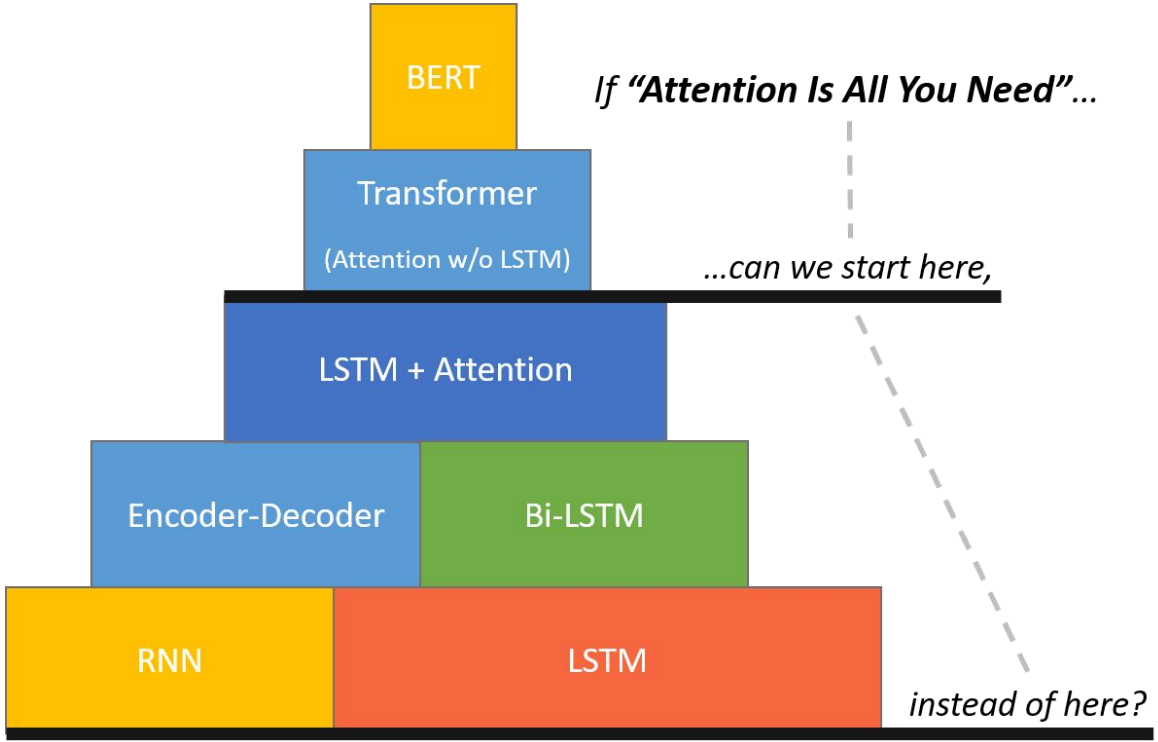



living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Introdução

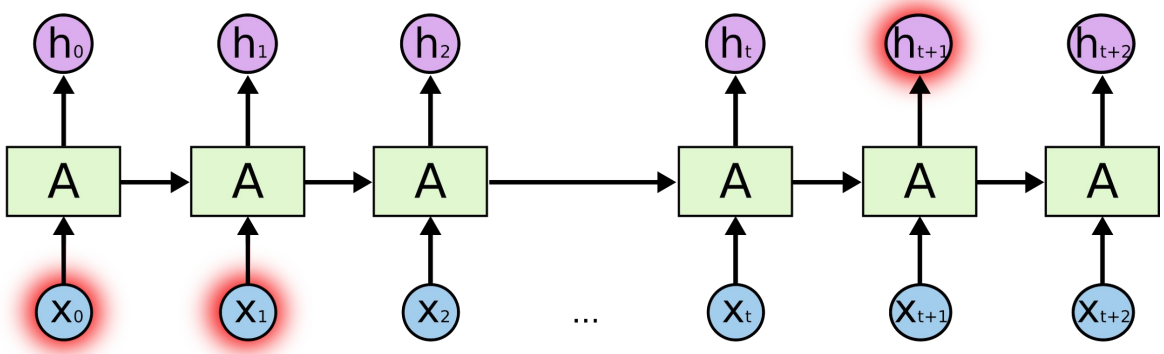
- Próximos passos: como interpretar o contexto das palavras?
  - Texto como um sinal no tempo (a ordem importa):
    - Uso de Redes Recorrentes.
  - Evolução: *Transformers*.

Introdução


BERT
Transformer
(Attention w/o LSTM)
LSTM + Attention
Encoder-Decoder
Bi-LSTM
RNN
LSTM
If "Attention Is All You Need"…
…can we start here,
instead of here?

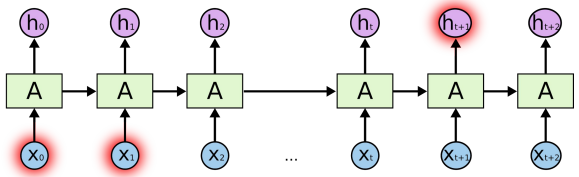
Introdução

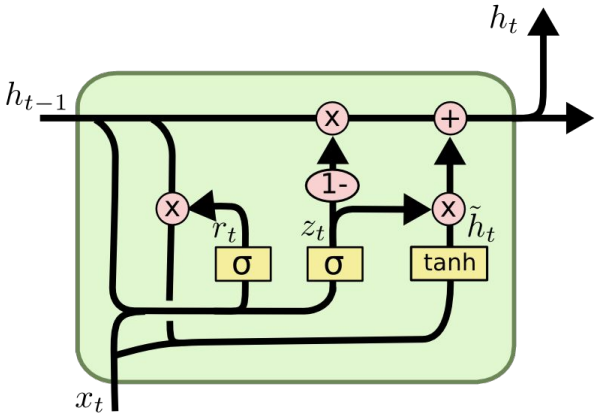
- Redes Recorrentes
  - *Long Short-Term Memory (LSTM)*
    - Os “neurônios” de uma camada conectam entre si de forma sequencial



Introdução

- *Long Short-Term Memory (LSTM)*





$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

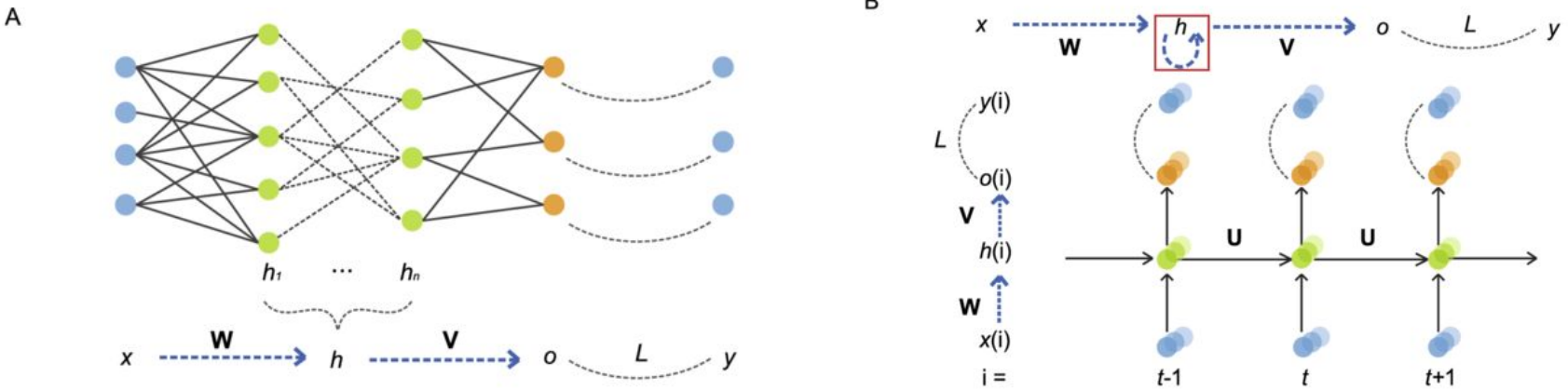
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

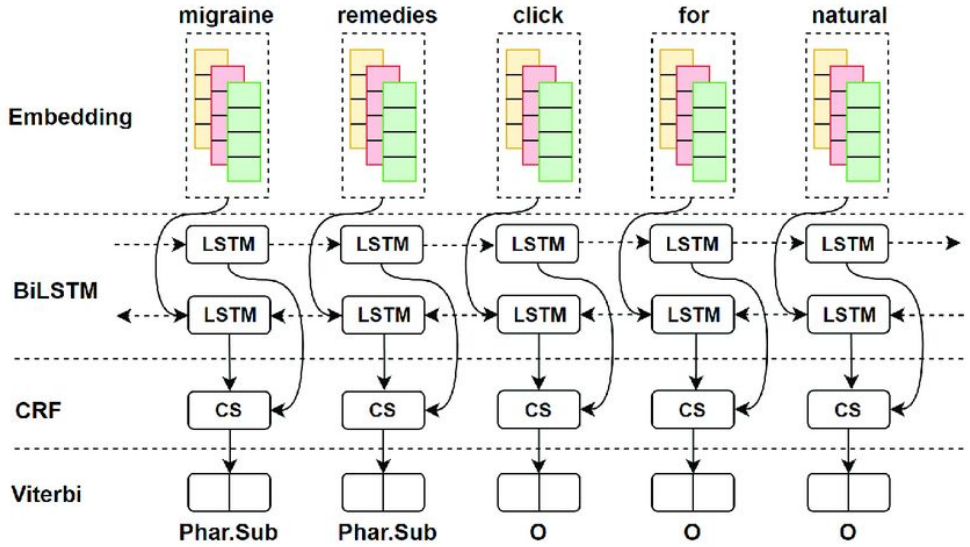
Introdução

- *Feedforward Network x Long Short-Term Memory (LSTM)*



Introdução

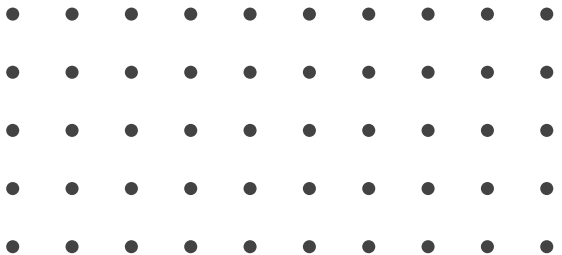
- *LSTM Bidirecional (BiLSTM)* - O futuro também influencia o presente



02

Transformers

"Atenção é tudo o que você precisa"



Transformers: o surgimento

- 7 pesquisadores do Google;
- Artigo "*Attention is All You Need*";
- Conferência "*Advances in Neural Information Processing Systems*", 2017;
- 13911 citações.

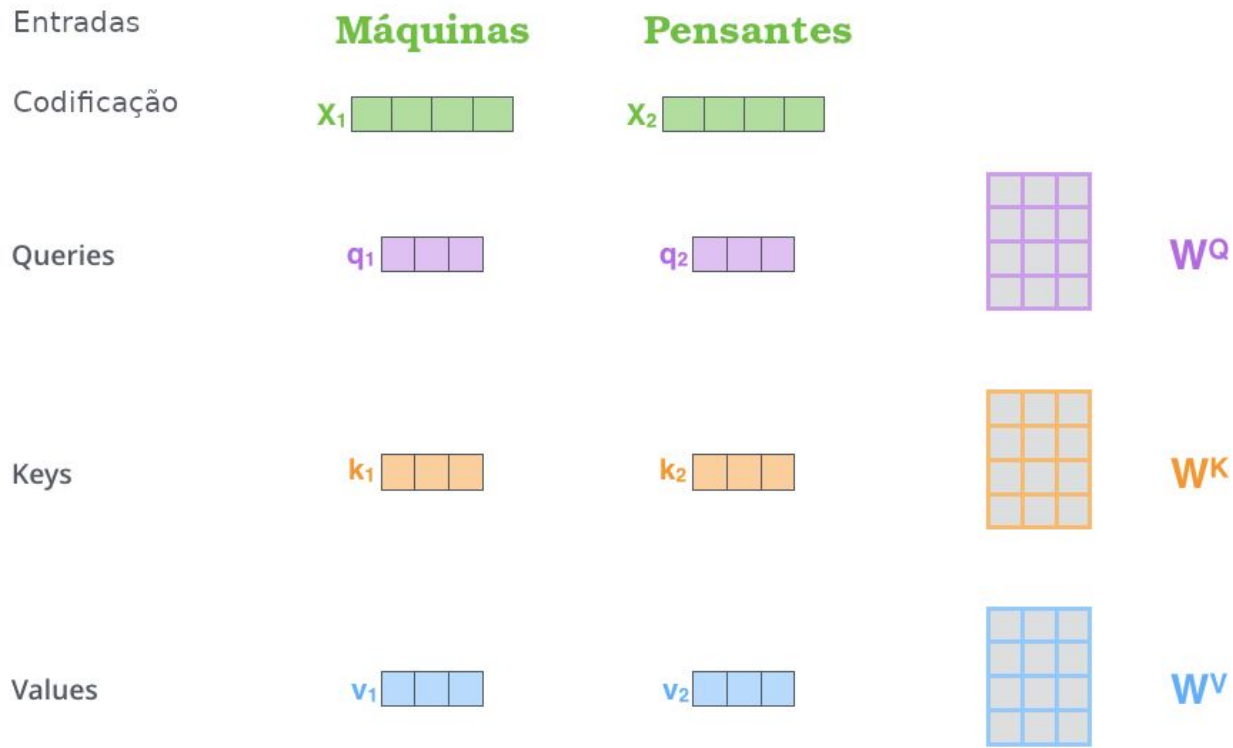
Motivação

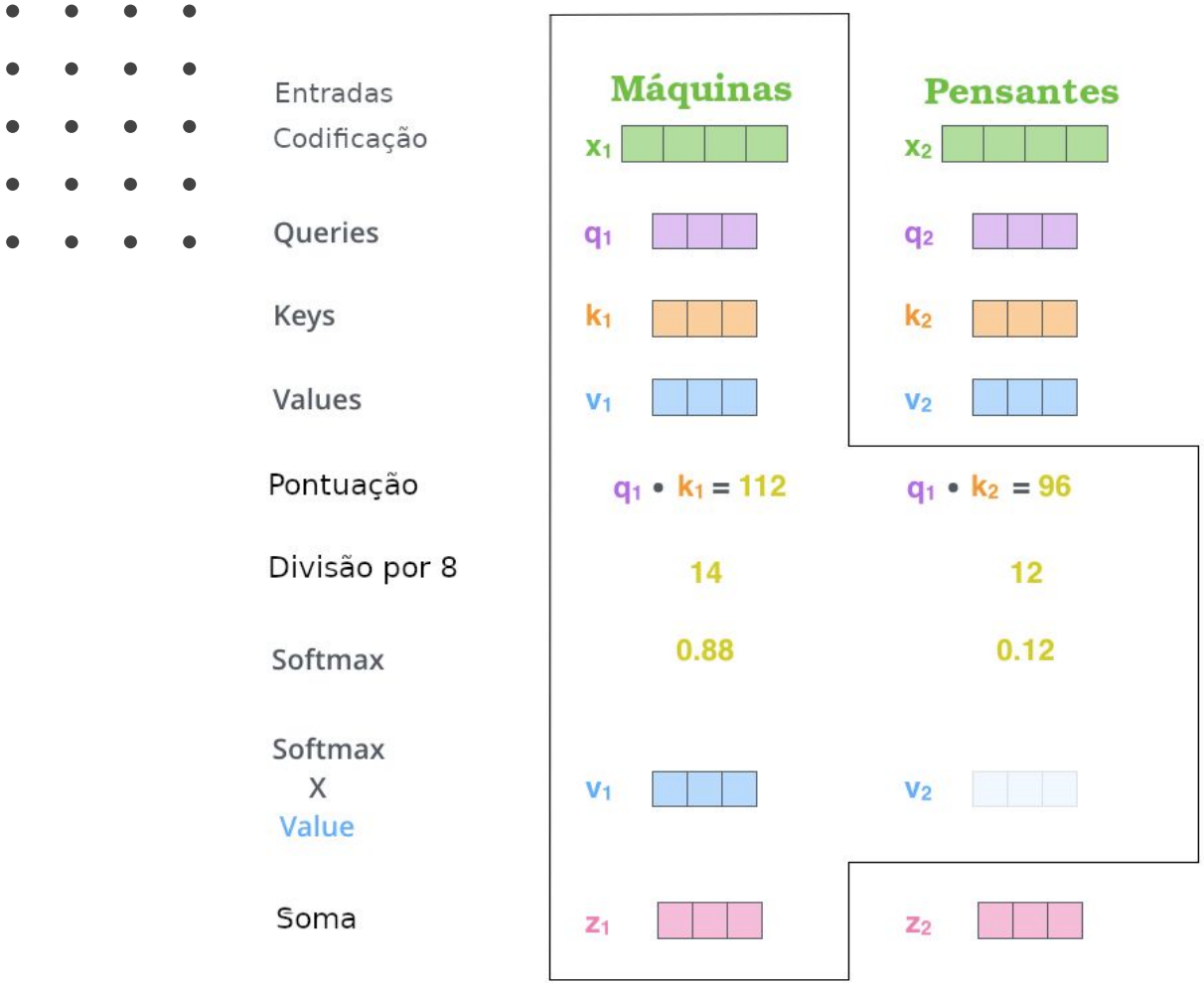
- Pré-*transformers:* redes recorrentes;
- Vantagens dos *transformers*:
  - Menor complexidade de computação por camada;
  - Mais propício à paralelização;
  - Menor espaço entre as dependências de palavras distantes

“ATENÇÃO”

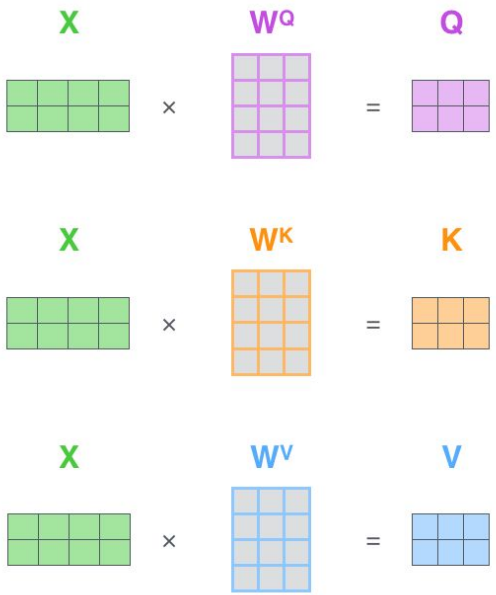
“O pássaro não voou sobre o rio porque ele estava cansado”

Atenção: o mecanismo




Entradas
Codificação
Máquinas
Pensantes
x_1
x_2
Queries
q_1
q_2
Keys
k_1
k_2
Values
v_1
v_2
Pontuação
$q_1 \cdot k_1 = 112$
$q_1 \cdot k_2 = 96$
Divisão por 8
14
12
Softmax
0.88
0.12
Softmax
X
Value
v_1
v_2
Soma
z_1
z_2

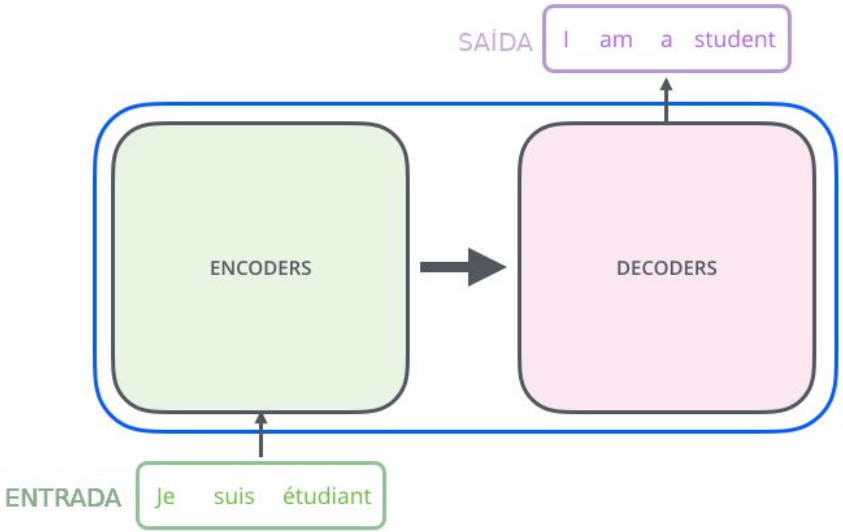
Atenção: o mecanismo



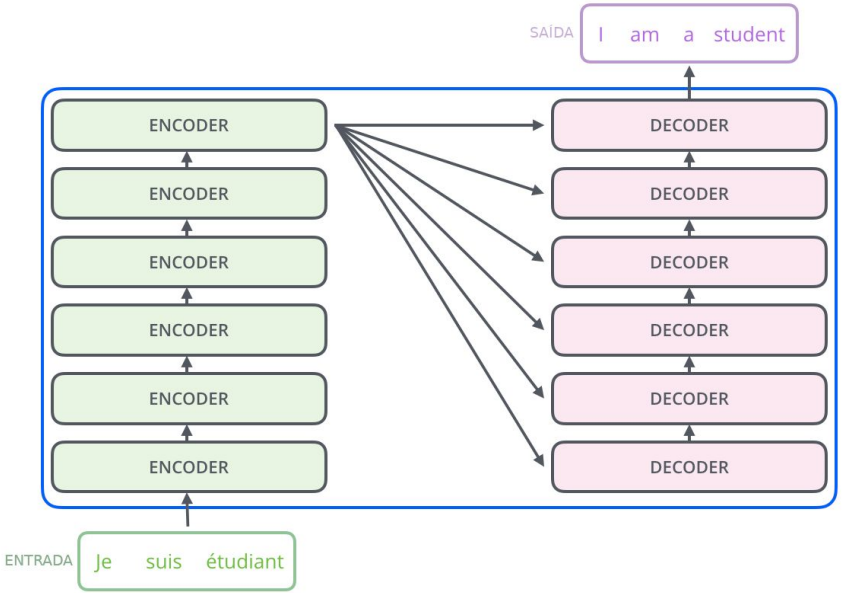
$$\text{softmax}\left(\frac{Q \times K^T}{8}\right) V = Z$$

Arquitetura dos *transformers*

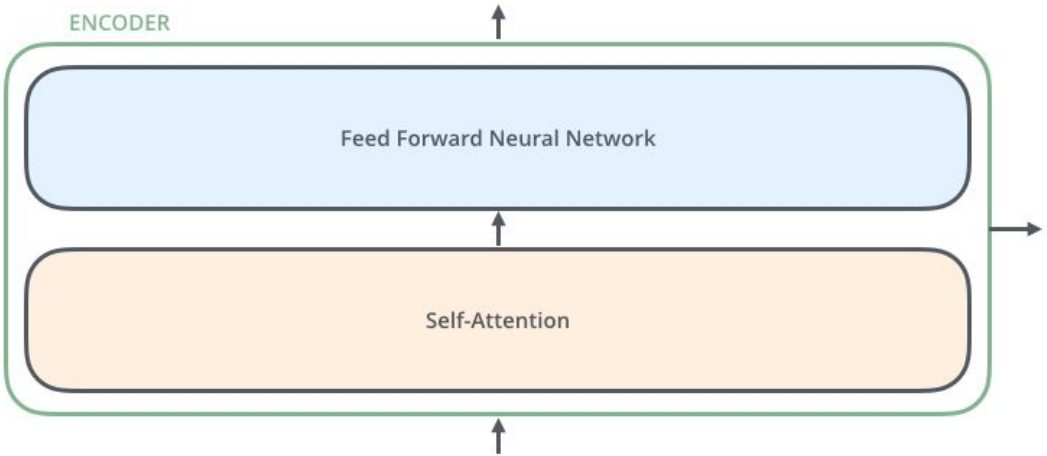
Arquitetura dos *transformers*



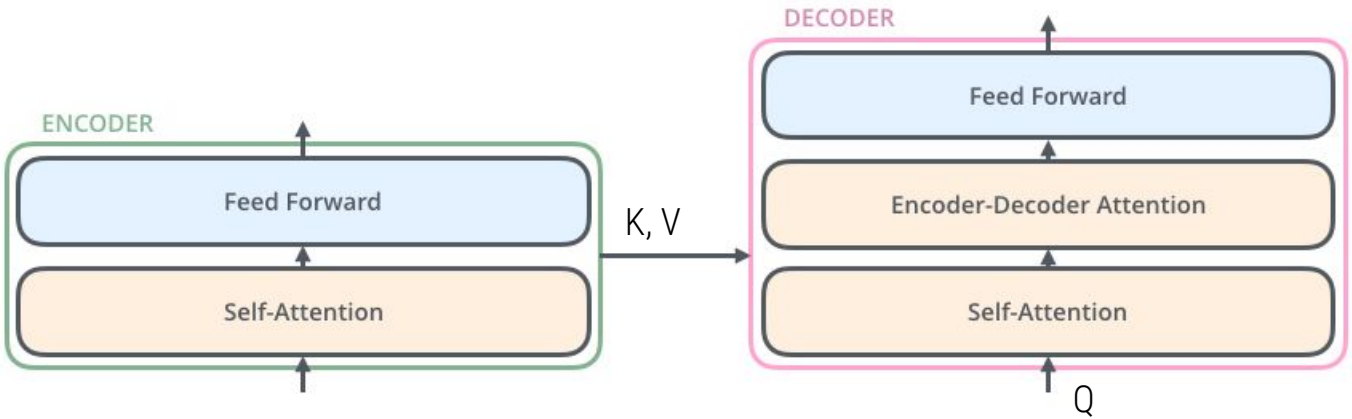
Arquitetura dos *transformers*



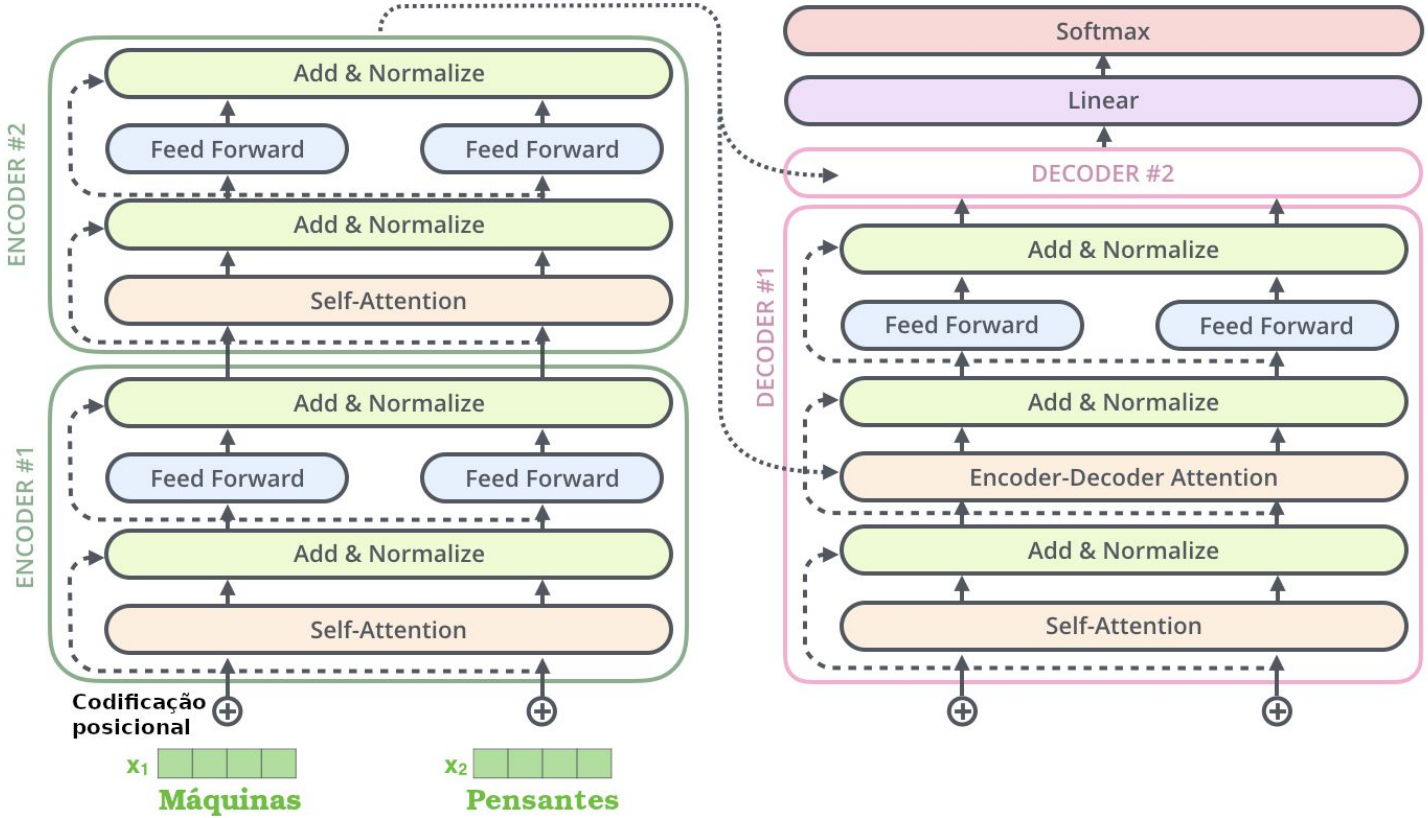
Arquitetura dos *transformers*



Arquitetura dos *transformers*



Visão geral


Softmax
Linear
DECODER #2
DECODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Encoder-Decoder Attention
Add & Normalize
Self-Attention
ENCODER #2
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
ENCODER #1
Add & Normalize
Feed Forward
Feed Forward
Add & Normalize
Self-Attention
Codificação posicional
x1
x2
Máquinas
Pensantes

Multi-headed attention

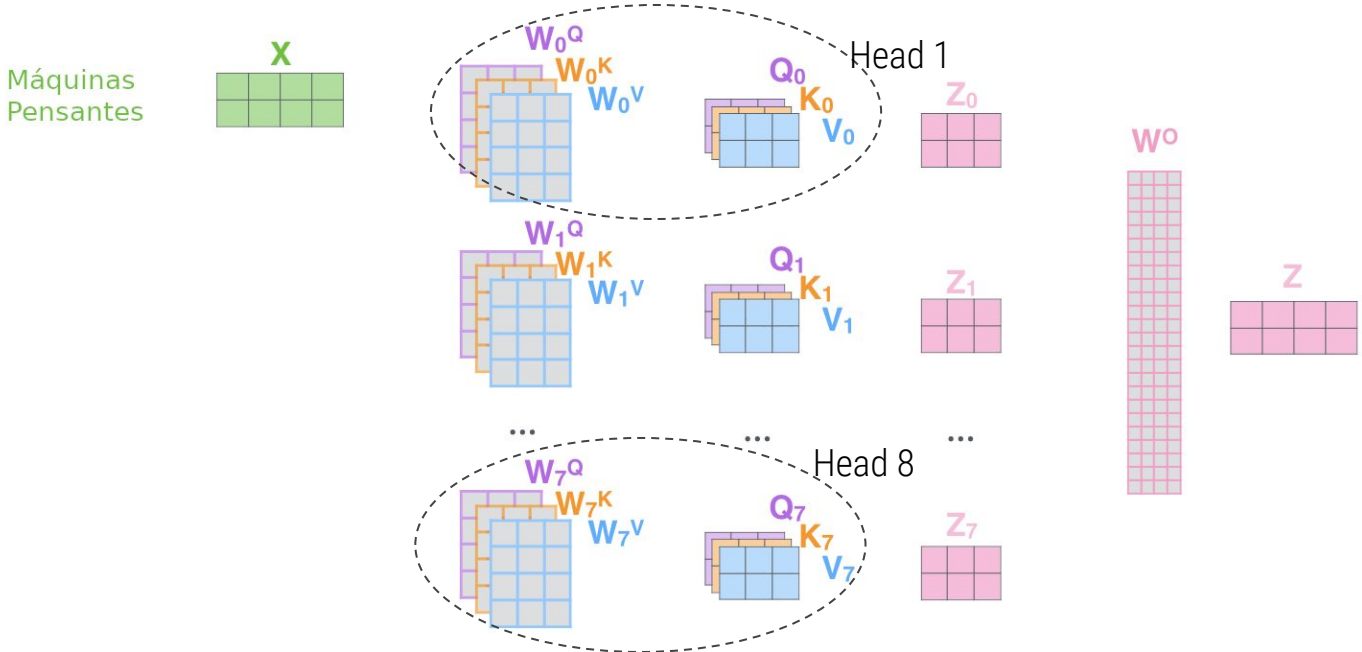
1) Entrada

2) Entrada codificada

3) Dividir em 8 cabeças e multiplicar X por cada uma delas.

4) Calcular atenção usando as matrizes **Q**/**K**/**V**

5) Concatenar as matrizes Z_i e multiplicar o resultado com uma matriz de pesos W^O para produzir a saída.



RESULTADOS

MODELO	BLEU		Custo de treino (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
GNMT+RL	24.6	39.92	2,3*10^19	1.4 *10^20
ConvS2S	25.16	40.46	9,6*10^18	1,5 * 10^20
MoE	26.03	40.56	2,0 * 10 ^19	1,2 * 10^20
Transformer	28.4	41.8	2,3 * 10^19	2,3 * 10^19

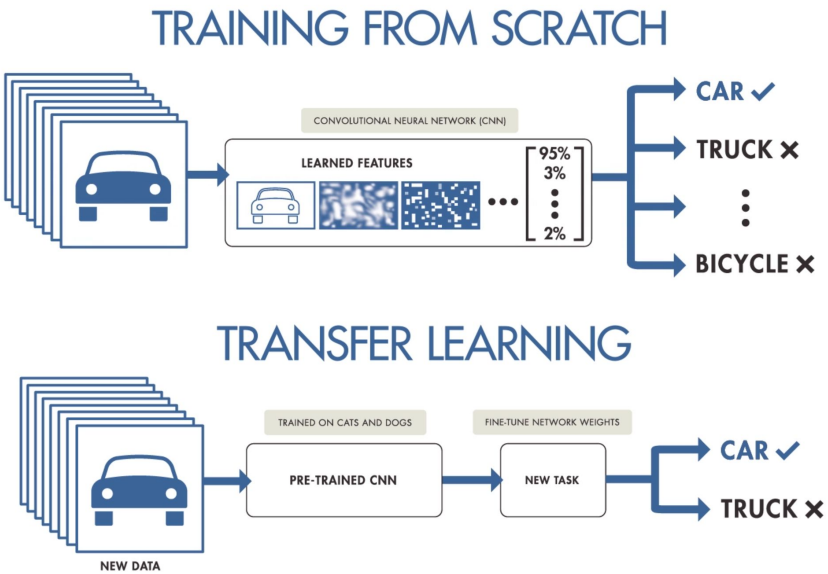
03

Google BERT

"Bidirectional Encoder Representations from Transformers"
Representação Bidirecional de Contexto

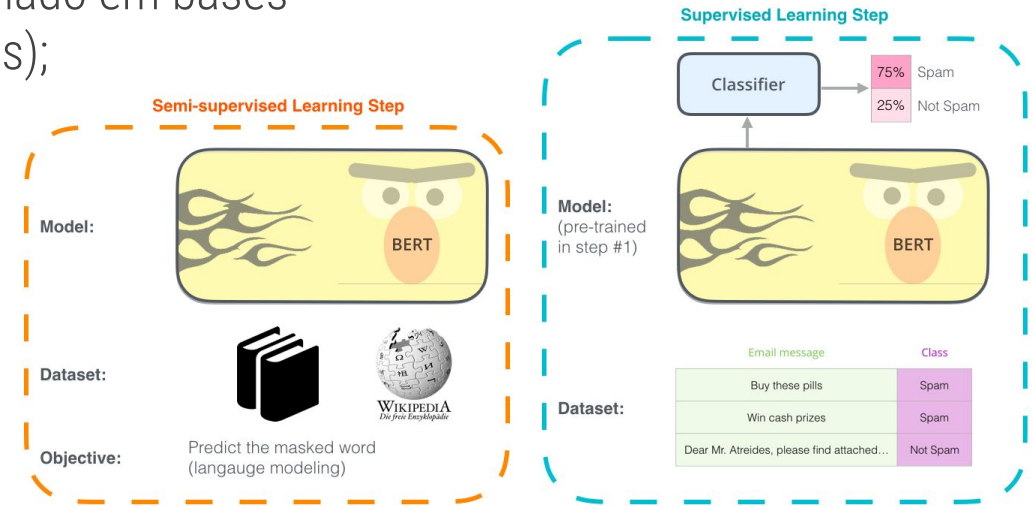
Técnica de *Transfer Learning*

- Redes treinadas com grande volume de dados e tarefas genéricas;
- Nova rede com os mesmos pesos da antiga e novas camadas antes da saída (fine-tuning layer);
- Novo treinamento -> ajusta os pesos.



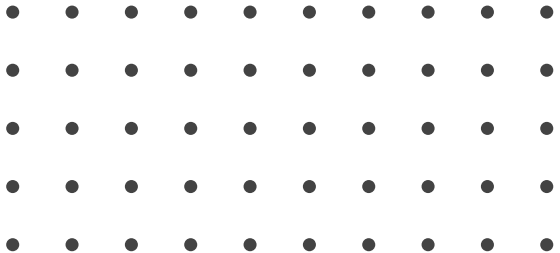
Técnica de *Transfer Learning*

- Modelos baseados em transformers usam aprendizado semi-supervisionado em bases gigantescas (bilhões de dados);
- *Fine-tuning* para as mais diversas tarefas possíveis;
- Por ter conhecimento prévio, supera a maioria das estruturas em resultado.



Google BERT

- Considerado um marco importante na história da IA;
- Artigo *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"* do Google AI:
  - Publicado em 2018 na NAACL;
  - Com cerca de 12 mil citações atualmente.
- Implementa os modelos de atenção e a arquitetura *encoder-decoder;*
- Foco principal: perguntas e respostas.



Google BERT

- Modelo de Linguagem Bidirecional

→ **Esquerda para Direita** →

"All of the BERT results presented so far have used the fine-tuning **[MASK]**"

← **Direita para Esquerda** ←

"**[MASK]** of the BERT results presented so far have used the fine-tuning approach"

↔ **Bidirecional** ↔

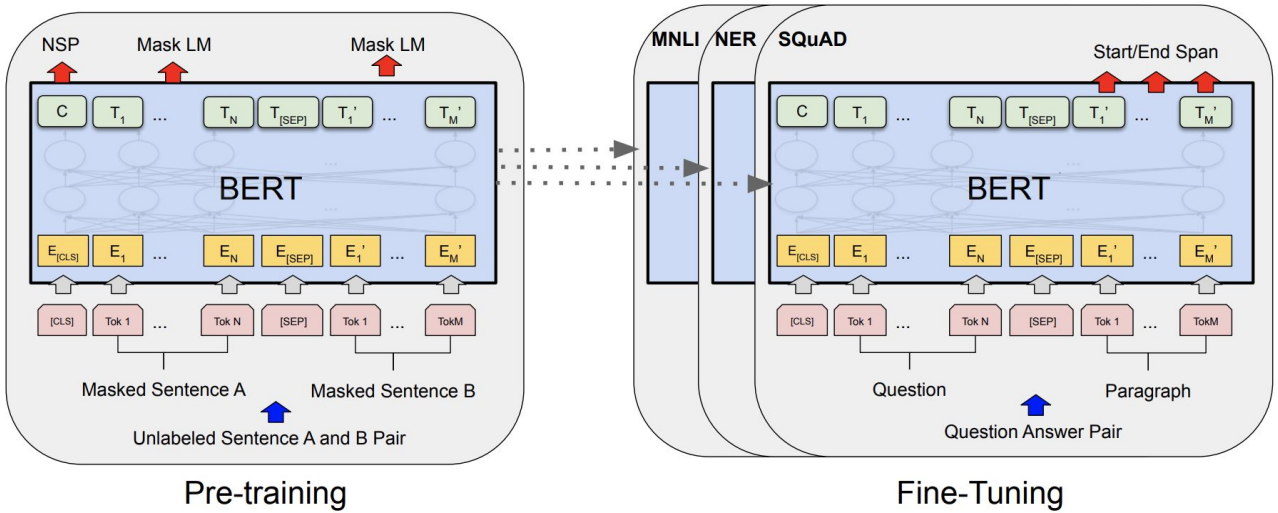
"All of the **[MASK]** results presented so **[MASK]** have used the fine-tuning approach"

Google BERT

- Estratégia de para auto-aprendizado:
  - Não precisa de dados anotados;
  - Esconde 15% das palavras da base;
  - Dados de Treino: Bilhões de textos - livros e wikipédia.
- 80% dos casos - esconde uma palavra:
  - *My dog is hairy -> My dog is [MASK];*
- 10% dos casos - troca a palavra por uma aleatória:
  - *My dog is hairy -> My dog is apple.*
- 10% dos casos - mantém a mesma palavra:
  - *My dog is hairy -> My dog is hairy.*

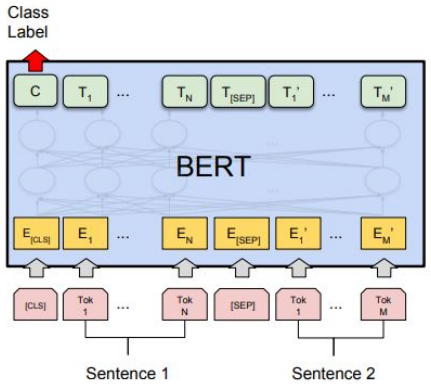
Google BERT

- Modelo geral

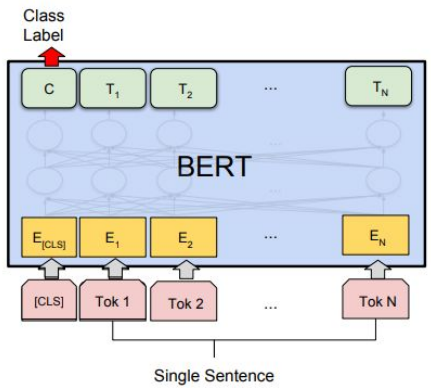


Google BERT

- *Fine-tuning*
  - Ajustar a arquitetura para a tarefa desejada.



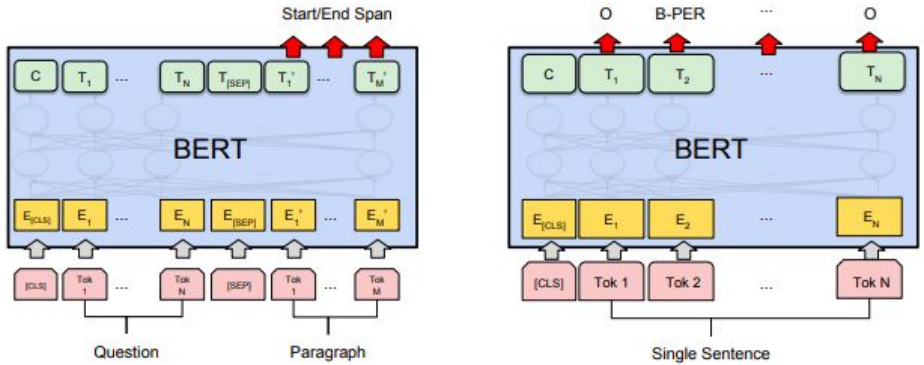
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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

Google BERT

- *Fine-tuning*
  - Ajustar a arquitetura para a tarefa desejada.

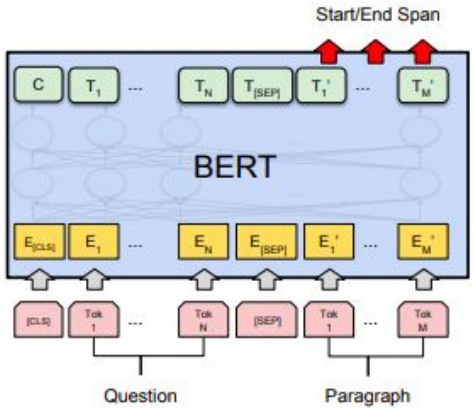


(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

Google BERT

- Exemplo: Pergunta e resposta.



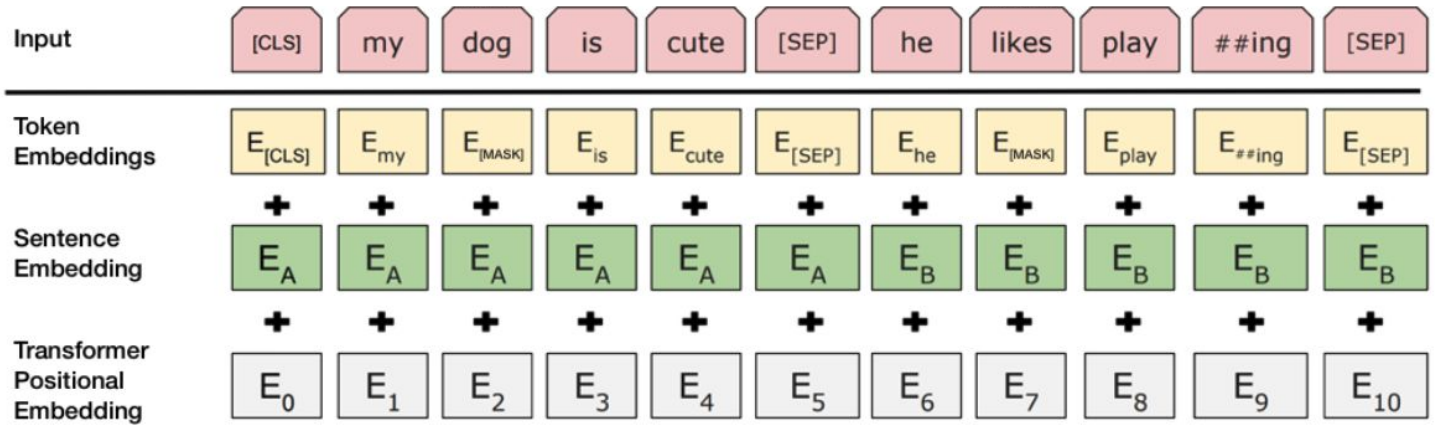
(c) Question Answering Tasks: SQuAD v1.1

When did the Paris Sevens become the last stop on the calendar?

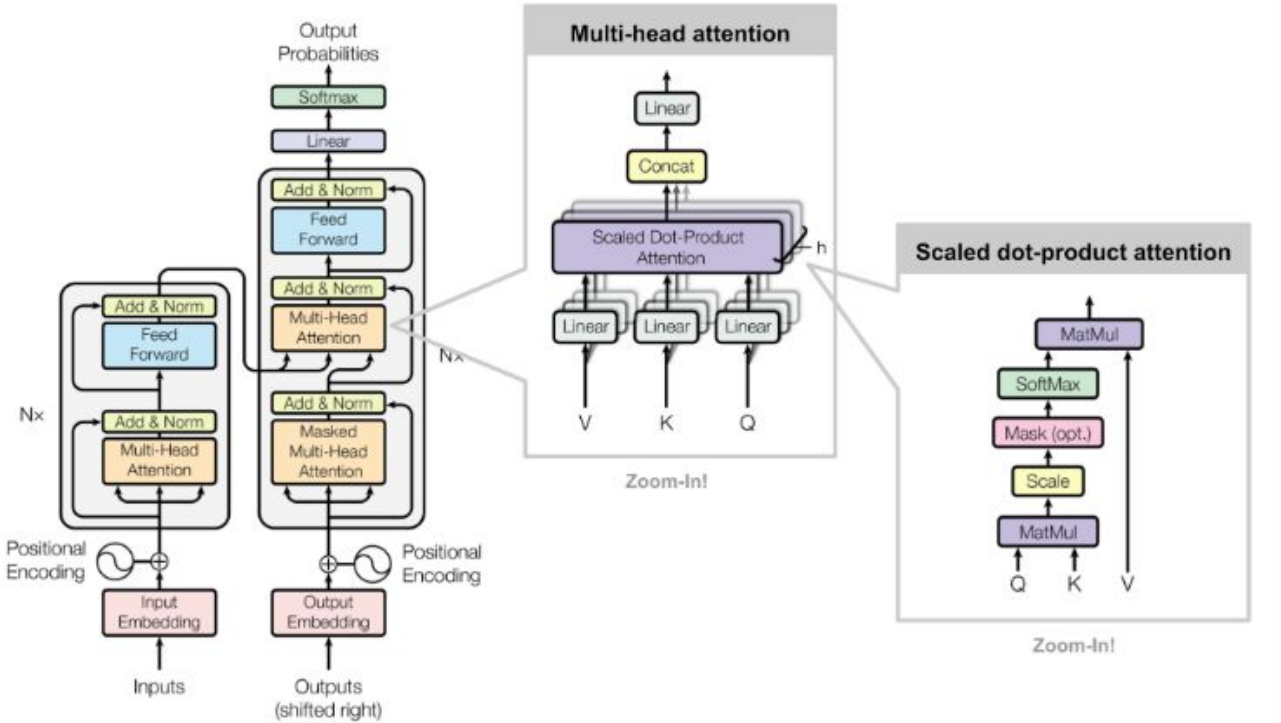
The London Sevens is a rugby tournament held at Twickenham Stadium in London. It is part of the World Rugby Sevens Series. For many years the London Sevens was the last tournament of each season but the Paris Sevens became the last stop on the calendar in 2018.

Google BERT

- Exemplo de Entrada
  - Tokenização - Codificar as palavras em números



Arquitetura BERT



Dicas BERT

- BERT é com certeza o *Transformer* com mais informações e códigos exemplos na internet, sendo muito fácil de treinar;
- Possui versão em diversas línguas, português inclusive;
- Encontra-se na TensorFlow Hub;
- Encontra-se na biblioteca *Transformers* da *HuggingFace* para *Keras/TF* ou *PyTorch*.

TensorFlow Hub

04

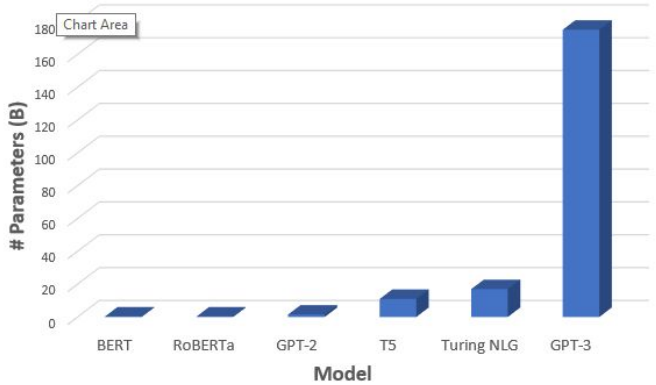
ESTADO DA ARTE

Um novo marco a cada mês.

GPT-3

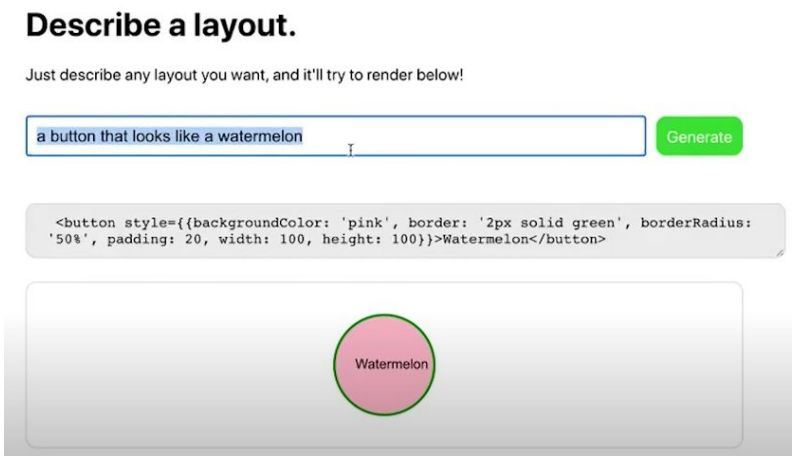
OpenAI

- GPT-3 é considerado o Estado da Arte em Transformers para Texto;
- Projeto da OpenAI anunciado em junho de 2020;
- 175 bilhões de parâmetros de treinamento.



GPT-3

- “Parece mágica”:
  - IA que produz IA;
  - IA que produz códigos em Python;
  - IA que produz códigos em React;
  - Pode generalizar uma tarefa com três dados de treinamento.
- Muito bom para tarefas muito sistemáticas;
- Mas na prática ele não entende o que está fazendo.



OUTROS TRANSFORMERS

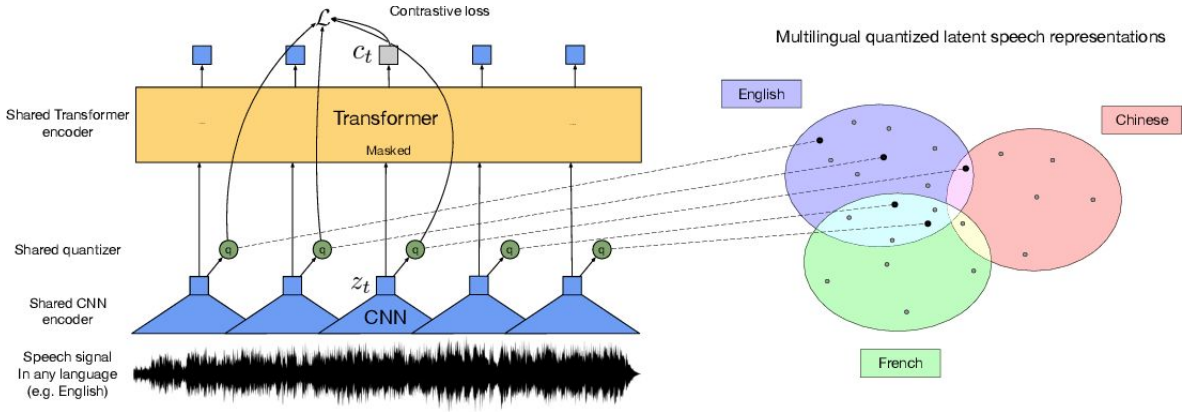
- BERTimbau (UNICAMP, 2020):
  - Estado da arte para o português.

- BERTweet (VinAI, 2020):
  - Supera qualquer outra rede em tarefas relacionadas a classificação e interpretação de tweets.

WAV2VEC

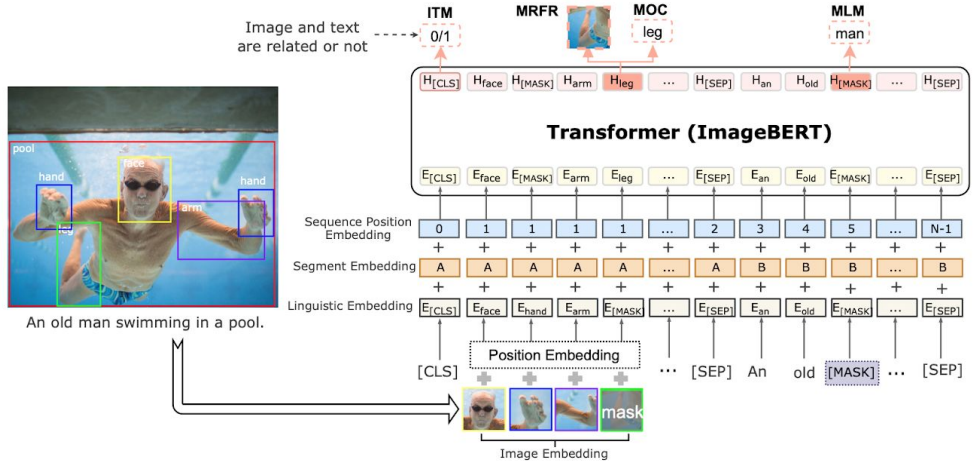
facebook research

- Facebook (2019) busca uma tecnologia similar aos transformers para identificar e classificar som;
- Uso de máscaras e modelos de atenção;
- Novas versões usam conceitos de Processamento de Texto.



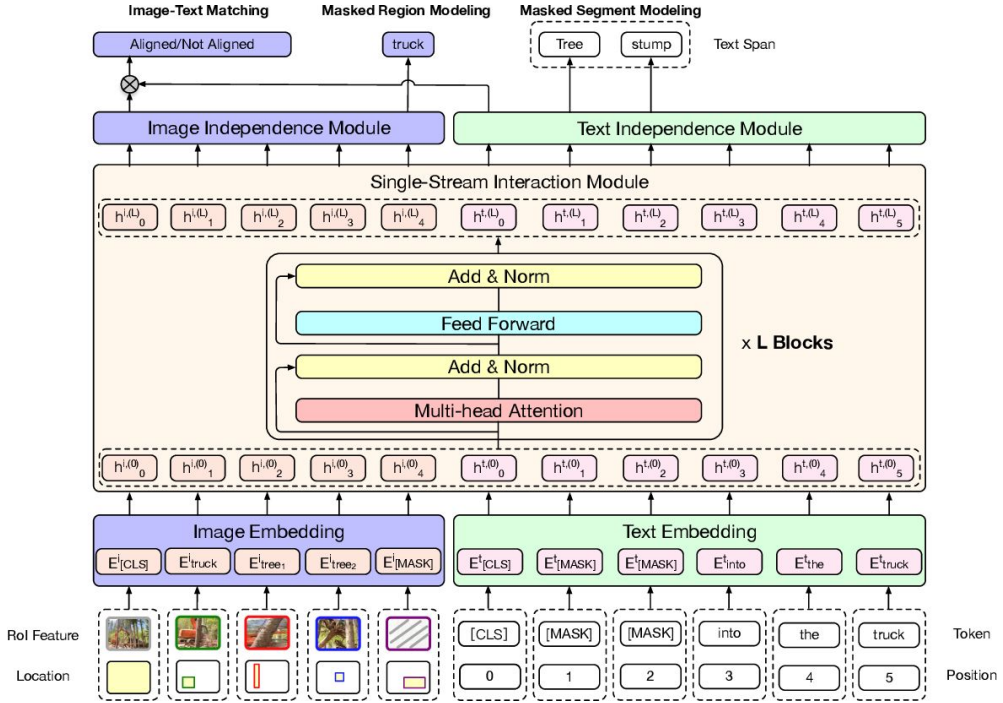
ImageBERT

- Projeto da Microsoft (Jan, 2020);
- Conceito de Image Embedding;
- Avanços rumo à sistemas multi-canais.



InterBERT

- Projeto da Alibaba (Jun, 2020) na mesma linha da Microsoft.



OBRIGADO!

Pedro Henrique Barbosa de Almeida, 10258793
Thomas Palmeira Ferraz, 9348985

REFERÊNCIAS

TRANSFORMERS

- VASWANI, Ashish et al. Attention is all you need. In: Advances in neural information processing systems. 2017. p. 5998-6008.
- ALAMMAR, Jay. The Illustrated Transformer. Disponível em: <https://jalammar.github.io/illustrated-transformer/>. Acesso em: 02/09/2020.
- CHAUBARD, Francois; SOCHER, Richard. Natural Language Processing with Deep Learning Lecture Notes. Disponível em: <http://web.stanford.edu/class/cs224n/>. Acesso em: 02/09/2020.

BERT

- DEVLIN, Jacob. Contextual Word Representations with BERT and Other Pre-trained Language Models. Disponível em: <http://web.stanford.edu/class/cs224n/slides/Jacob_Devlin_BERT.pdf>. Acesso em: 02/09/2020.
- DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Disponível em: <https://arxiv.org/pdf/1810.04805.pdf>. Acesso em: 02/09/2020.
- SOUZA, Fabio et al. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. Disponível em: <https://link.springer.com/chapter/10.1007%2F978-3-030-61377-8_28>. Acesso em: 02/09/2020.
- NGUYEN, Dat Quoc et al. BERTweet: A pre-trained language model for English Tweets. Disponível em: <https://arxiv.org/pdf/2005.10200.pdf>. Acesso em: 02/09/2020.
- LIN, Junyang. InterBERT: An Effective Multi-Modal Pretraining Approach via Vision-and-Language Interaction. Disponível em: <https://arxiv.org/pdf/2003.13198.pdf>. Acesso em: 02/09/2020.