Pontificia Universidad JAVERIANA Colombia

Procesamiento de Lenguaje Natural

Tópicos Avanzados en Analítica

Maestría en Analítica para la Inteligencia de Negocios

Sergio Alberto Mora Pardo - H2 2023

BERT:
Bidirectional Encoder Representations from Transformers

BERT

Bidirectional Encoder Representation for Transformers (2018)

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. BERT alleviates the previously mentioned unidirectionality constraint by using a "masked language model" (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Google

- La predicción del modelo Masked-language (Mask LM). (15%)

- La predicción de la siguiente oración (NSP). (50%)

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Masked-Language

El modelo BERT toma como entrada dos oraciones separadas por un token especial ([SEP])

Input: the man went to the [MASK1] . he bought a [MASK2] of milk.

Labels:
[MASK1] = store
[MASK2] = gallon

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Next Sentence Prediction

El modelo BERT toma como entrada dos oraciones separadas por un token especial ([SEP])

Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .

Label: **IsNextSentence**

Sentence A: the man went to the store .
Sentence B: penguins are flightless.

Label: **NotNextSentence**

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

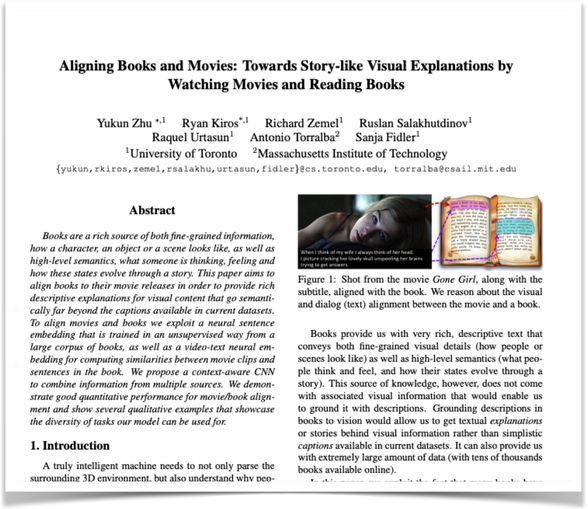
Función de perdida:

- *Mask ML*
- *NSP*

Wikipedia

+

BookCorpus



Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books

Yukun Zhu *,1 Ryan Kiros*,1 Richard Zemel[1] Ruslan Salakhutdinov[1]
Raquel Urtasun[1] Antonio Torralba[2] Sanja Fidler[1]
[1]University of Toronto [2]Massachusetts Institute of Technology
{yukun,rkiros,zemel,rsalakhu,urtasun,fidler}@cs.toronto.edu, torralba@csail.mit.edu

Abstract

Books are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story. This paper aims to align books to their movie releases in order to provide rich descriptive explanations for visual content that go semantically far beyond the captions available in current datasets. To align movies and books we exploit a neural sentence embedding that is trained in an unsupervised way from a large corpus of books, as well as a video-text neural embedding for computing similarities between movie clips and sentences in the book. We propose a context-aware CNN to combine information from multiple sources. We demonstrate good quantitative performance for movie/book alignment and show several qualitative examples that showcase the diversity of tasks our model can be used for.

1. Introduction

A truly intelligent machine needs to not only parse the surroundine 3D environment. but also understand whv peo-

Figure 1: Shot from the movie *Gone Girl*, along with the subtitle, aligned with the book. We reason about the visual and dialog (text) alignment between the movie and a book.

Books provide us with very rich, descriptive text that conveys both fine-grained visual details (how people or scenes look like) as well as high-level semantics (what people think and feel, and how their states evolve through a story). This source of knowledge, however, does not come with associated visual information that would enable us to ground it with descriptions. Grounding descriptions in books to vision would allow us to get textual *explanations* or stories behind visual information rather than simplistic *captions* available in current datasets. It can also provide us with extremely large amount of data (with tens of thousands books available online).

Alinear libros y películas: hacia explicaciones visuales parecidas a historias viendo películas y leyendo libros

BERT

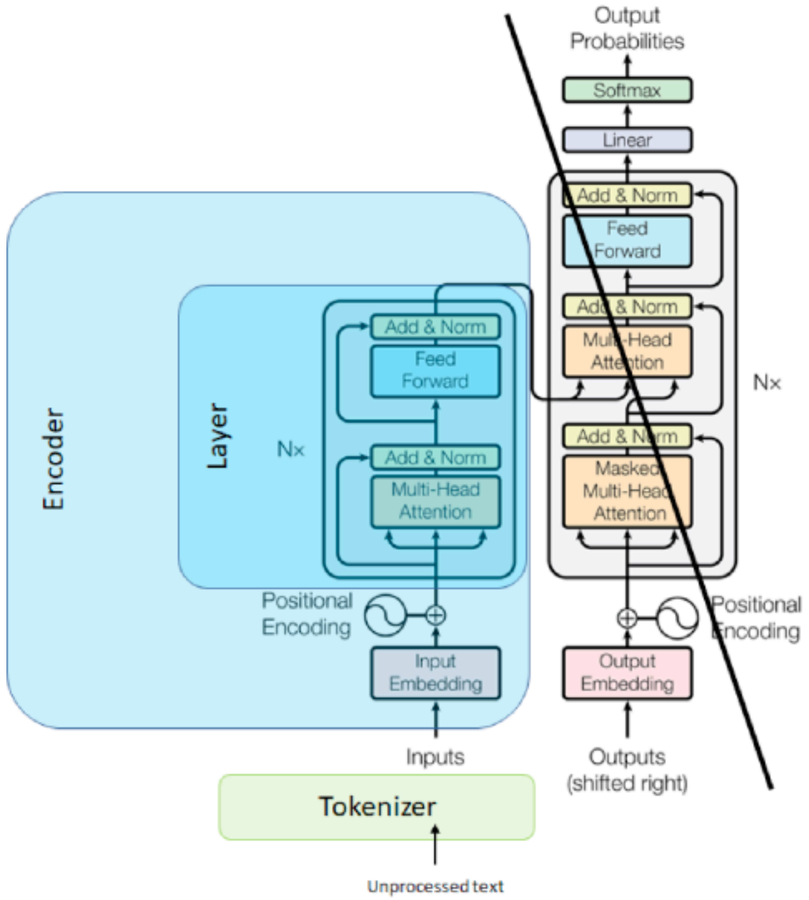
Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*



BERT

Bidirectional Encoder Representation for Transformers

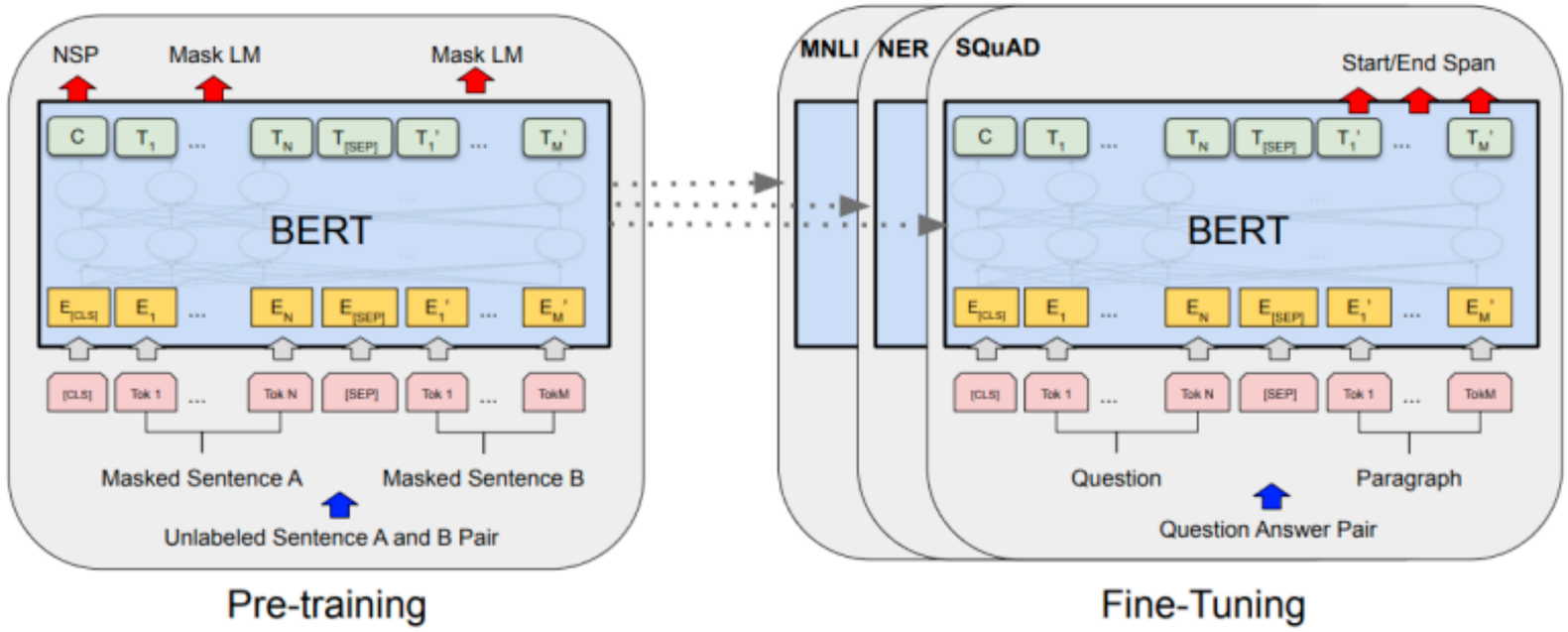
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT es utilizado a menudo como un modelo *encoder* de lenguaje.



Puede entenderse como un punto de chequeo (checkpoint) pre-entrenado el cual es extendido con capas adicionales que son específicas a la tarea de interés

BERT

Bidirectional Encoder Representation for Transformers

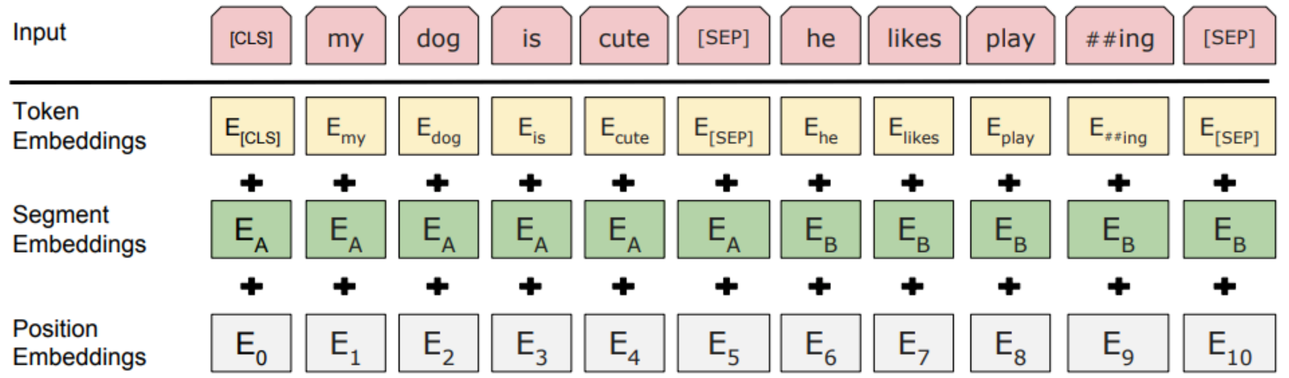
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Input representation



Usa los tokens [SEP] y [CLS] para cada oración y [MASK] para pronosticar esa misma palabra.

BERT

Bidirectional Encoder Representation for Transformers

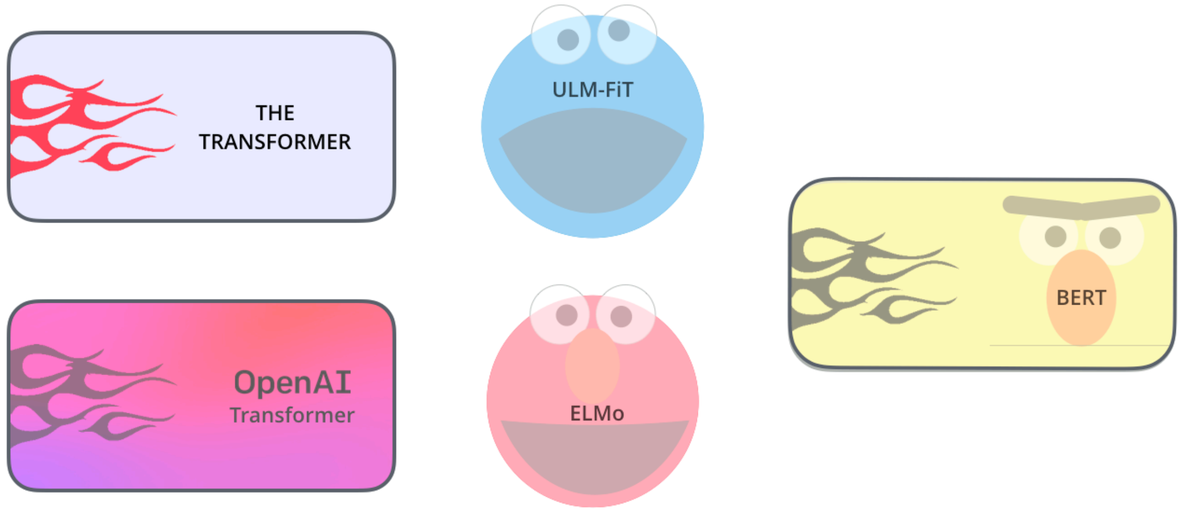
BERT

Pre-training of Deep
Bidirectional Transformers
for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT family... (RoBERT, RoBERTA, alBERT, ... Hasta un tal BETO existe)



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

BERT

Bidirectional Encoder Representation for Transformers

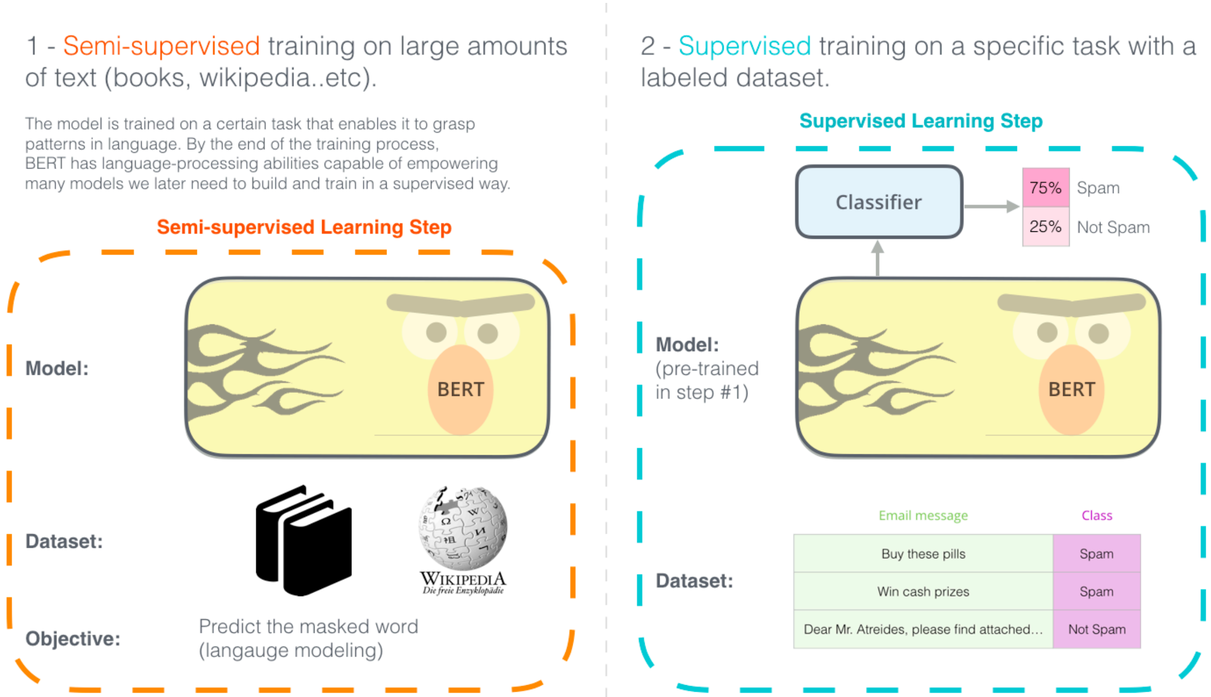
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Steps:



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

BERT

Bidirectional Encoder Representation for Transformers

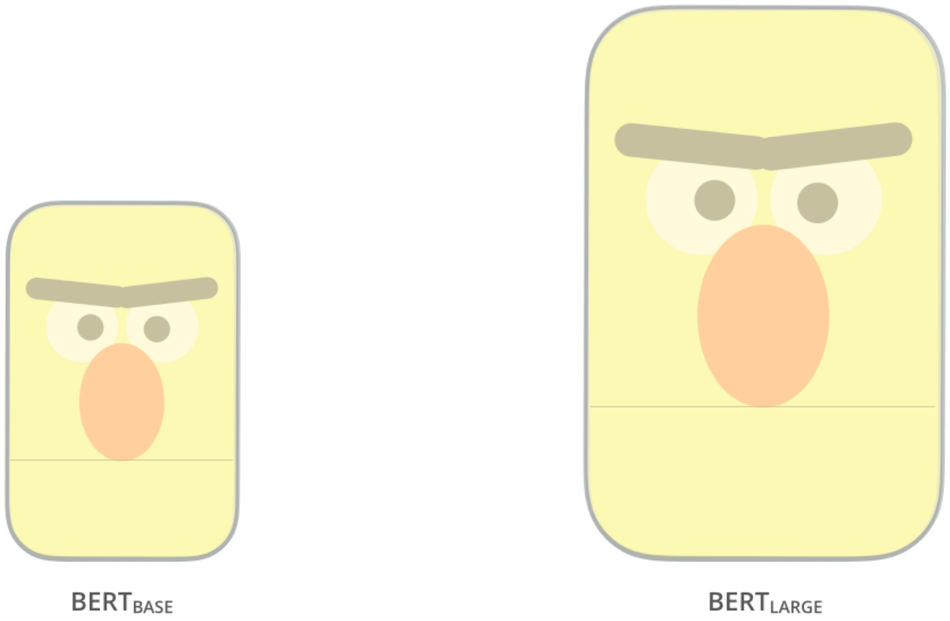
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Model arquitectura:



BERT-base de tamaño similar al OpenAI Transformer

BERT

Bidirectional Encoder Representation for Transformers

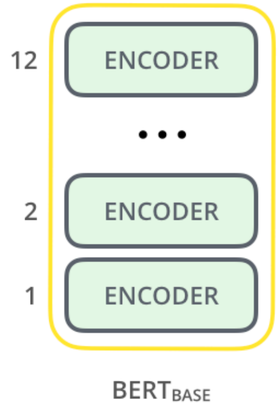
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

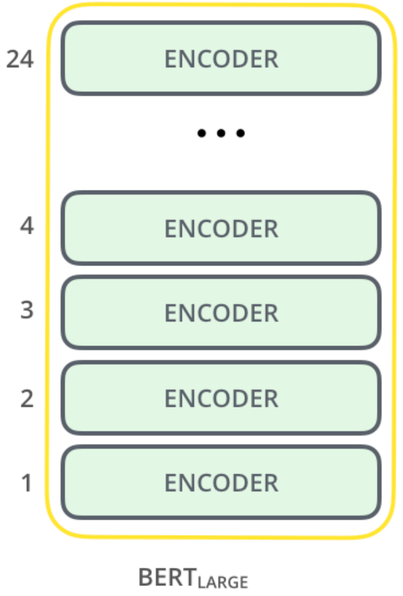
Model arquitectura:



$BERT_{BASE}$

768 dimensiones

12 cabezas de atención



$BERT_{LARGE}$

1024 unidades ocultas (dimensiones)

16 cabezas de atención

Transformer:

6 capas de encoder

512 dimensiones

8 cabezas de atención

BERT

Bidirectional Encoder Representation for Transformers

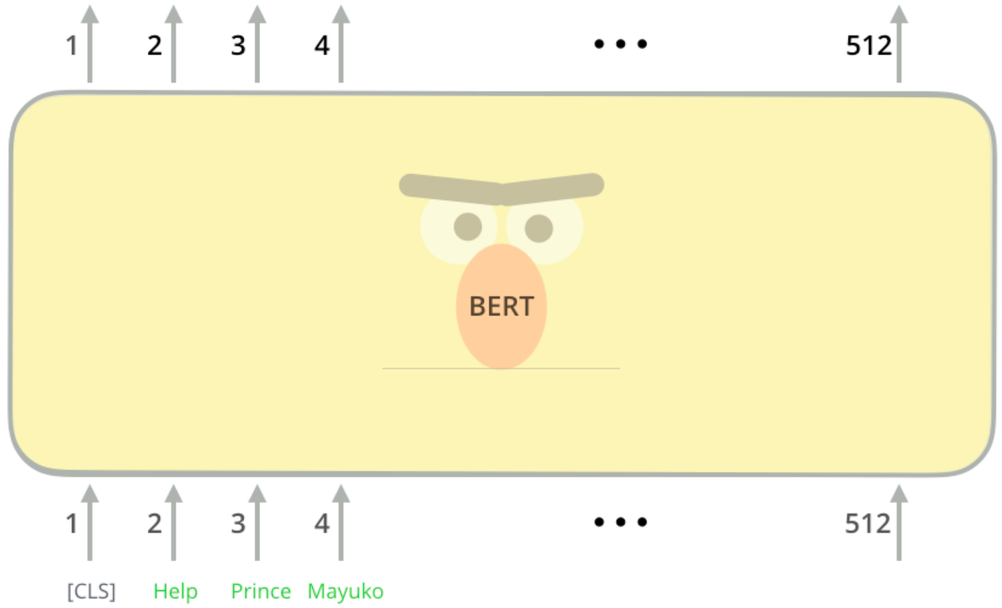
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Entradas del modelo



El primer token es [CLS] significa clasificación

BERT

Bidirectional Encoder Representation for Transformers

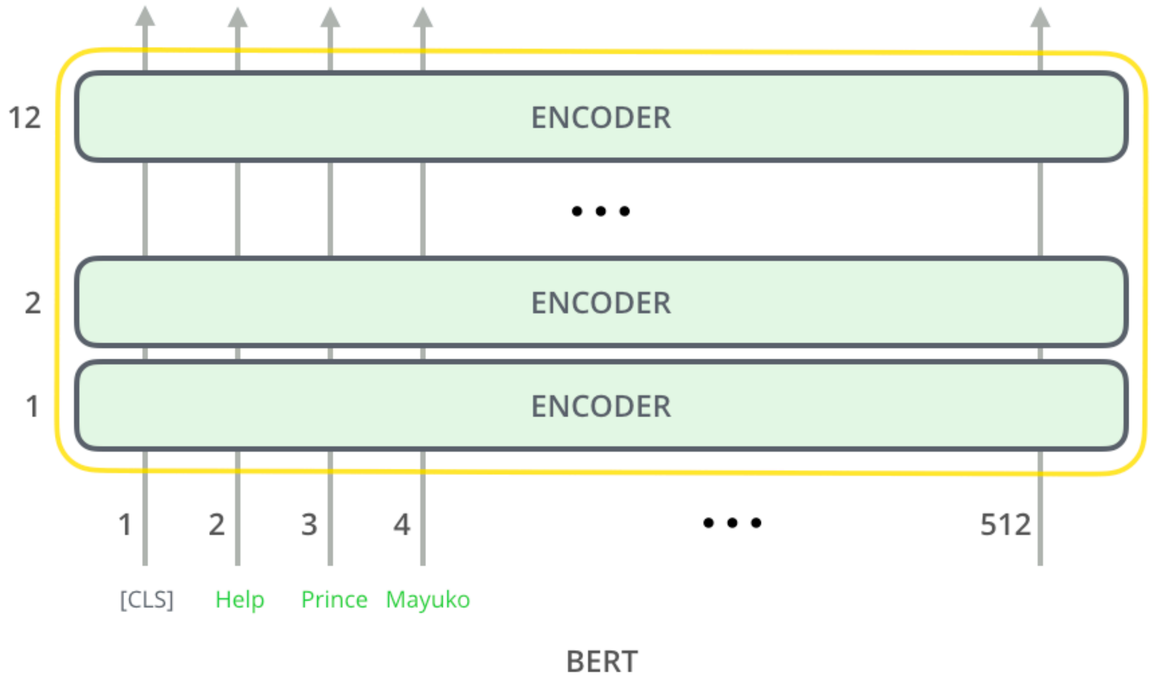
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Entradas del modelo



La arquitectura es idéntica al Transformer (hasta ahora)

BERT

Bidirectional Encoder Representation for Transformers

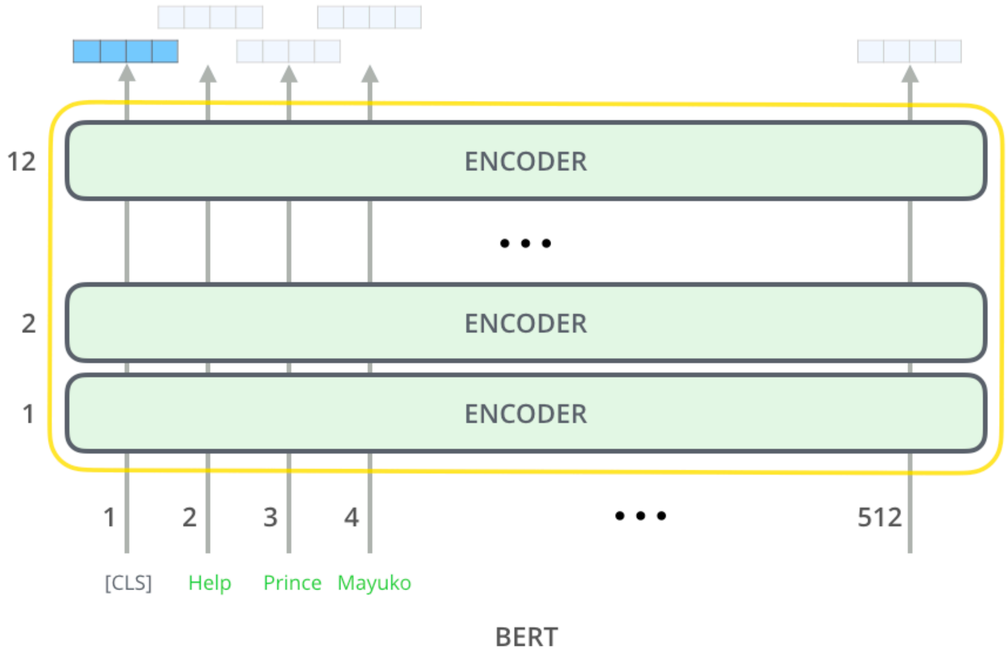
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Resultados del modelo



Cada posición genera un vector de tamaño 768 (BERT base)

Nos centramos en el vector con el token especial [CLS]

BERT

Bidirectional Encoder Representation for Transformers

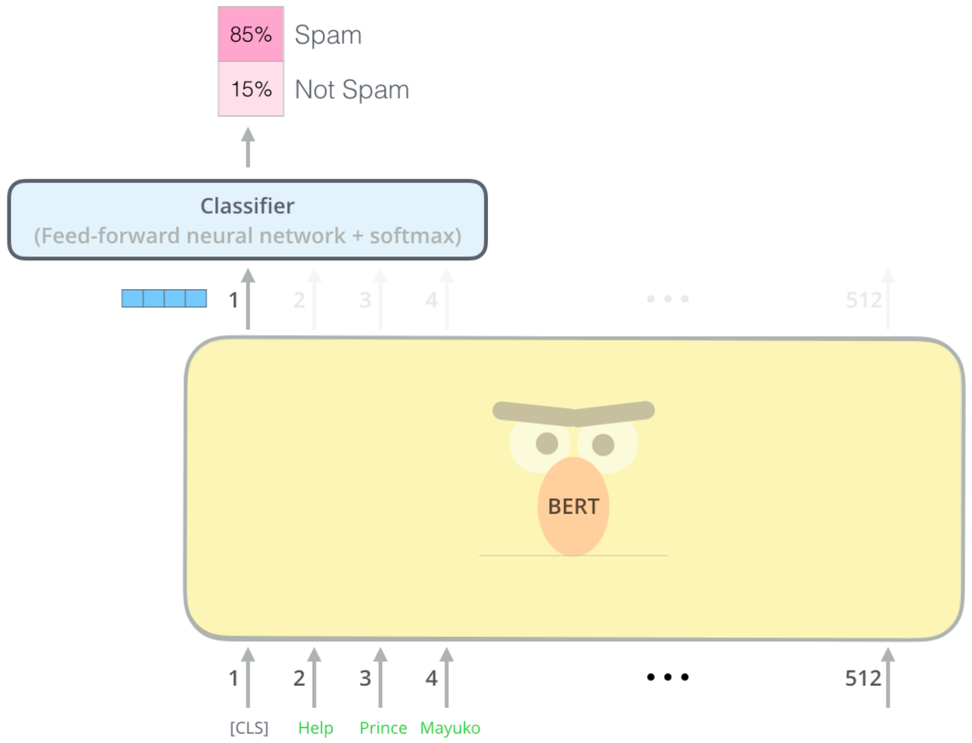
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Resultados del modelo



modelo de calificación

Ej.: spam-social-promoción

Ej.: spam

vector como entrada

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Resultados del modelo

modelo de calificación

Ej.: spam-social-promoción

Ej.: spam

vector como entrada

BERT

Bidirectional Encoder Representation for Transformers

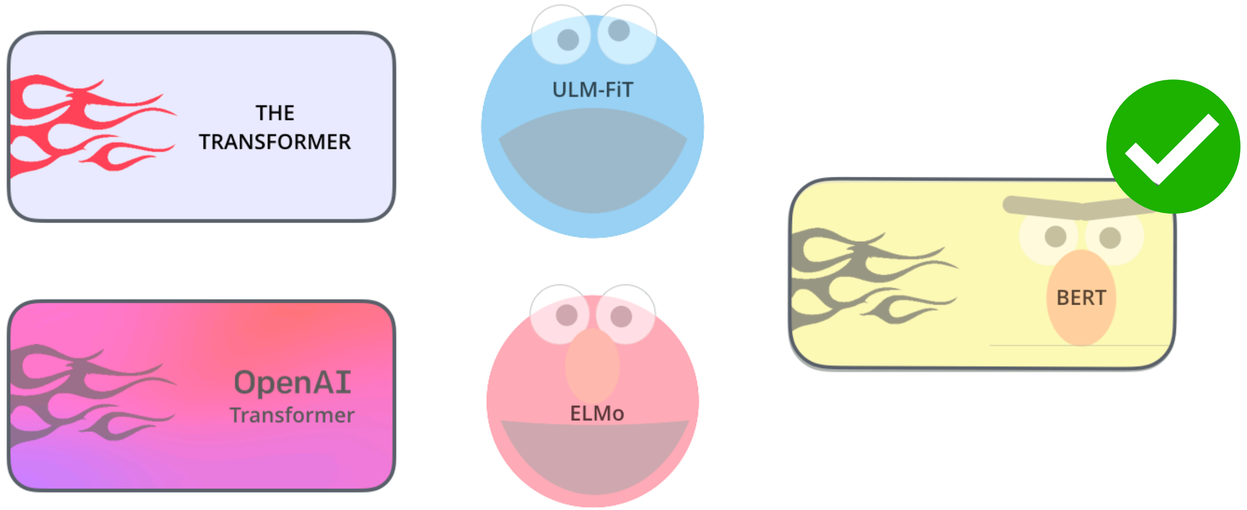
BERT

Pre-training of Deep
Bidirectional Transformers
for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT family... (RoBERT, RoBERTA, alBERT, ... Hasta un tal BETO existe)



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

ELMo

Embeddings from Language Models

ELMo

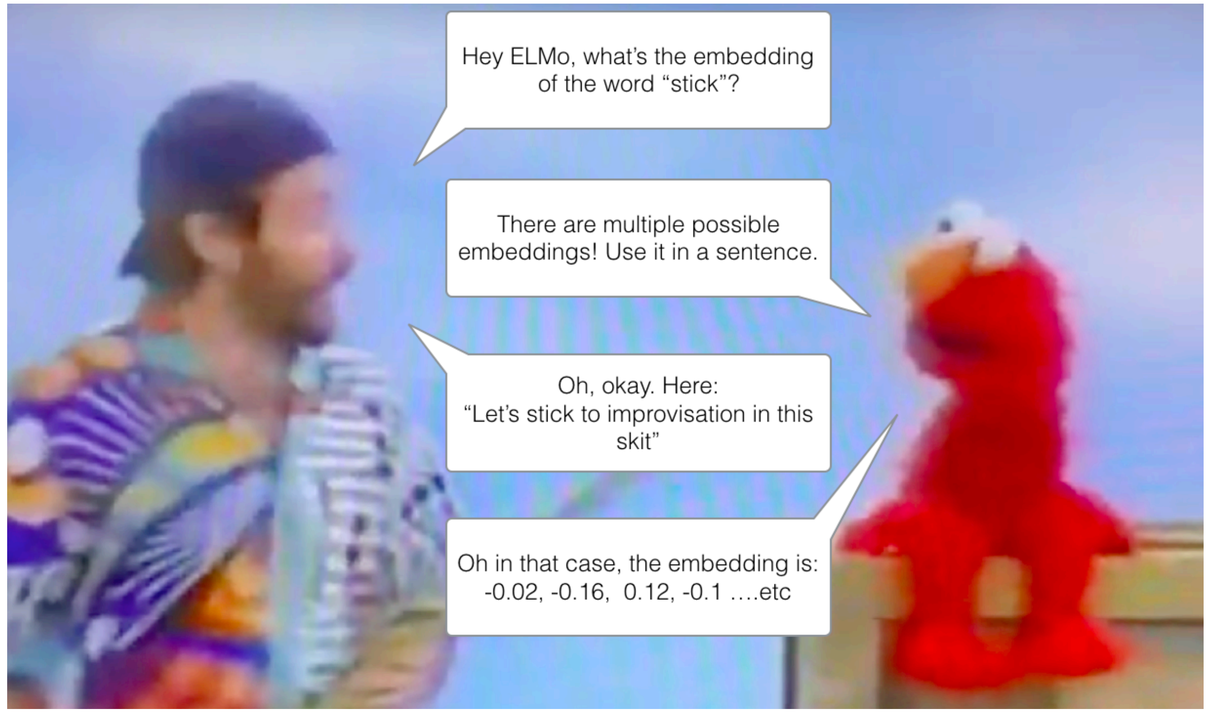
Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Context Matters



Representación semantica

Word2Vec *Glove*

↓

Medir distancia/similitud

Compresión en un espacio vectorial

ELMo

Embeddings from Language Models

ELMo

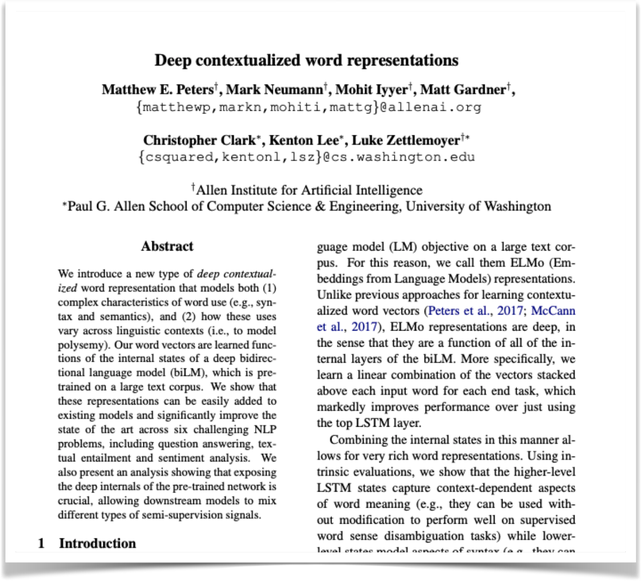
Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Context Matters



Deep contextualized word representations

Matthew E. Peters[†], Mark Neumann[†], Mohit Iyyer[†], Matt Gardner[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], Kenton Lee[*], Luke Zettlemoyer[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

Abstract

We introduce a new type of *deep contextualized* word representation that models both (1) complex characteristics of word use (e.g., syntax and semantics), and (2) how these uses vary across linguistic contexts (i.e., to model polysemy). Our word vectors are learned functions of the internal states of a deep bidirectional language model (biLM), which is pre-trained on a large text corpus. We show that these representations can be easily added to existing models and significantly improve the state of the art across six challenging NLP problems, including question answering, textual entailment and sentiment analysis. We also present an analysis showing that exposing the deep internals of the pre-trained network is crucial, allowing downstream models to mix different types of semi-supervision signals.

guage model (LM) objective on a large text corpus. For this reason, we call them ELMo (Embeddings from Language Models) representations. Unlike previous approaches for learning contextualized word vectors (Peters et al., 2017; McCann et al., 2017), ELMo representations are deep, in the sense that they are a function of all of the internal layers of the biLM. More specifically, we learn a linear combination of the vectors stacked above each input word for each end task, which markedly improves performance over just using the top LSTM layer.

Combining the internal states in this manner allows for very rich word representations. Using intrinsic evaluations, we show that the higher-level LSTM states capture context-dependent aspects of word meaning (e.g., they can be used without modification to perform well on supervised word sense disambiguation tasks) while lower-

1 Introduction

Peters et. al., 2018 en el artículo de ELMo

Representación semantica

Word2Vec *Glove*

↓

Medir distancia/similitud

Compresión en un espacio vectorial

ELMo

Embeddings from Language Models

ELMo

Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Context Matters

Representación semantica

Word2Vec *Glove*

↓

Medir distancia/similitud

Compresión en un espacio vectorial

Representación contextual

Capturar significado de la palabra en donde se usa

Agrega otra información contextual...

Peters et. al., 2018 en el artículo de ELMo

ELMo

Embeddings from Language Models

ELMo

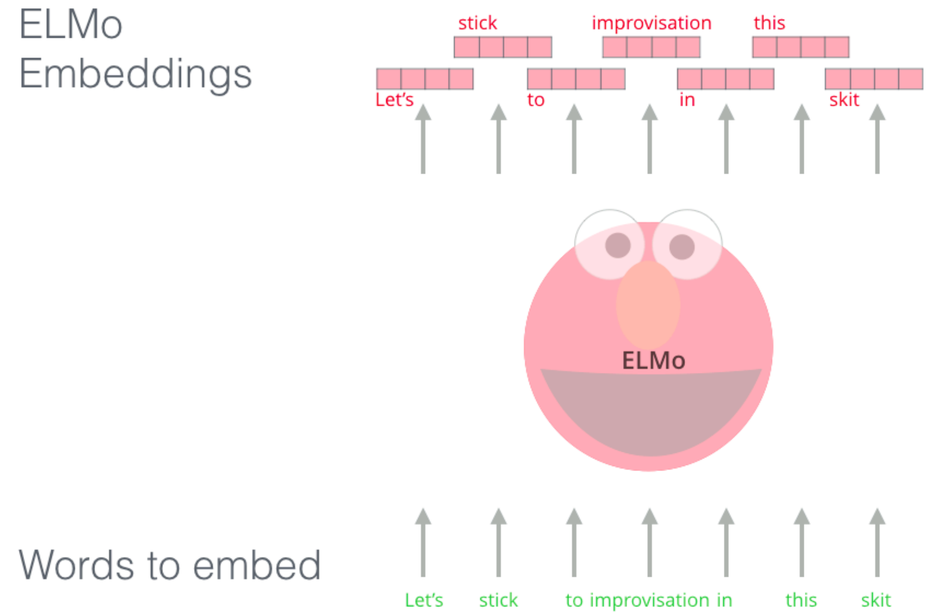
Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Context Matters



Representación contextual

Capturar significado de la palabra en donde se usa

Peters et. al., 2018 en el artículo de ELMo

ELMo

Embeddings from Language Models

ELMo

Embeddings from Language Models

LSTM bidireccional

↓

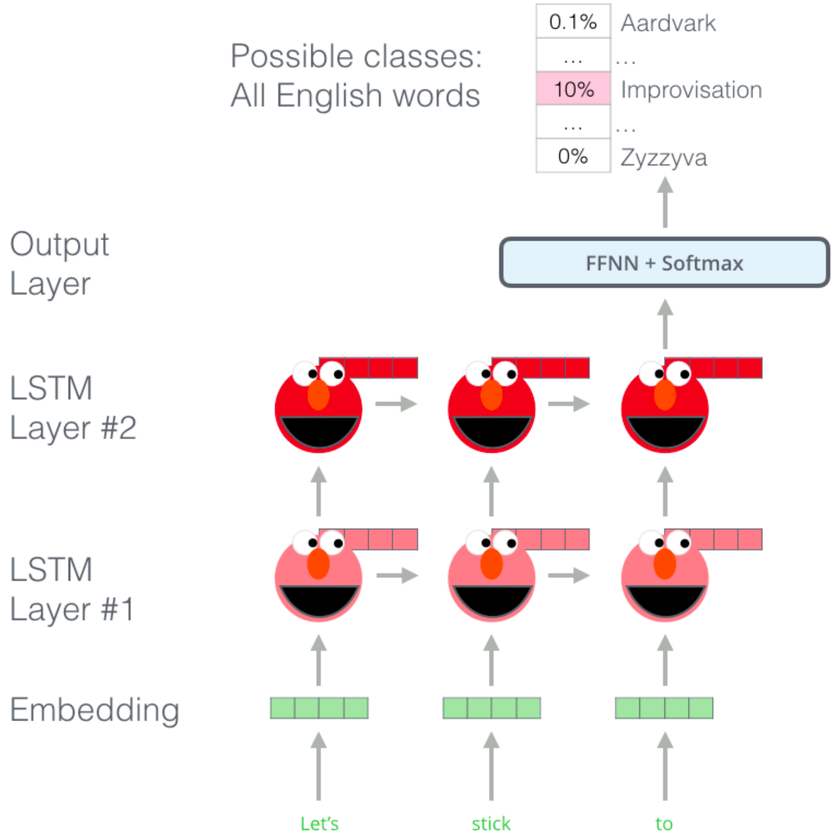
Revisa la oración completa

Context Matters

Modelo de Lenguaje

↓

Pronostica la siguiente palabra



Peters et. al., 2018 en el artículo de ELMo

ELMo

Embeddings from Language Models

ELMo

Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Context Matters

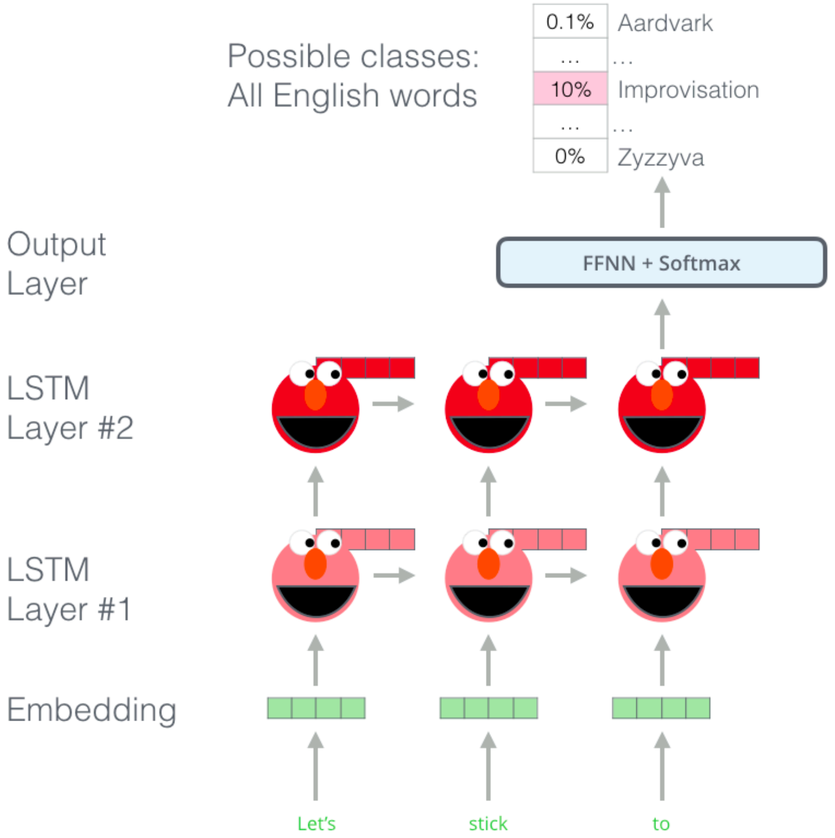
Modelo de Lenguaje

↓

Pronostica la siguiente palabra

+

Pronostica la palabra anterior



Peters et. al., 2018 en el artículo de ELMo

ELMo

Embeddings from Language Models

ELMo

Embeddings from Language Models

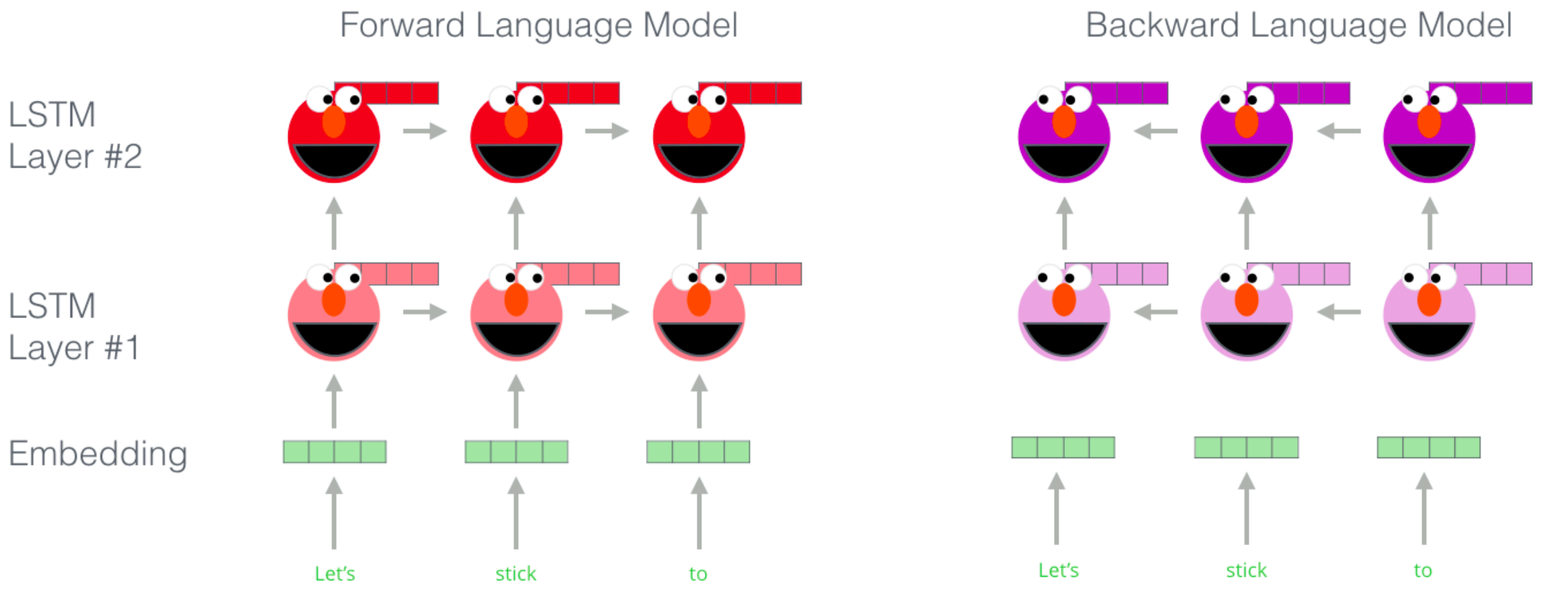
LSTM bidireccional

↓

Revisa la oración completa

Entrenamiento

Embedding of “stick” in “Let’s stick to” - Step #1



Peters et. al., 2018 en el artículo de ELMo

ELMo

Embeddings from Language Models

ELMo

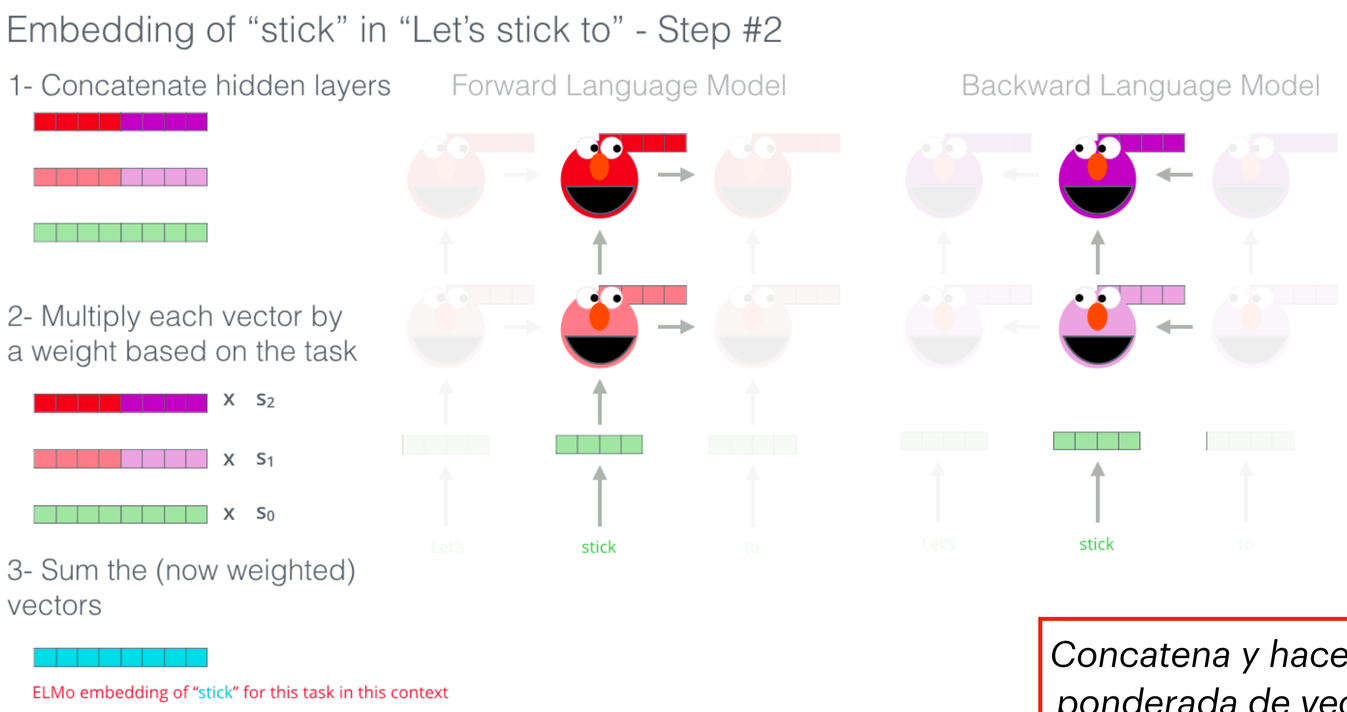
Embeddings from Language Models

LSTM bidireccional

↓

Revisa la oración completa

Entrenamiento



Peters et. al., 2018 en el artículo de ELMo

Concatena y hace suma ponderada de vectores

BERT

Bidirectional Encoder Representation for Transformers

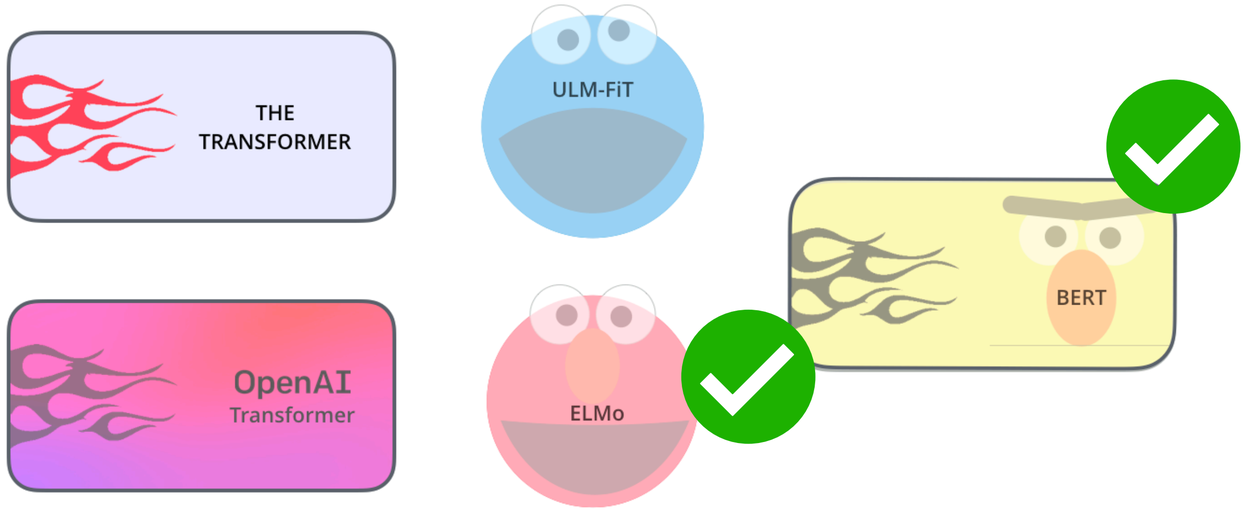
BERT

Pre-training of Deep
Bidirectional Transformers
for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT family... (RoBERT, RoBERTA, alBERT, ... Hasta un tal BETO existe)



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

BERT

Bidirectional Encoder Representation for Transformers

Estructura

Encoder-Decoder

Maneja mejor dependencias a largo plazo que los LSTM

Reemplazaría los LSTM en algunos campos

Modelo de lenguaje

Proceso para ajustar el modelo a diversas tareas

Introdujo el aprendizaje por transferencia

BERT

Bidirectional Encoder Representation for Transformers

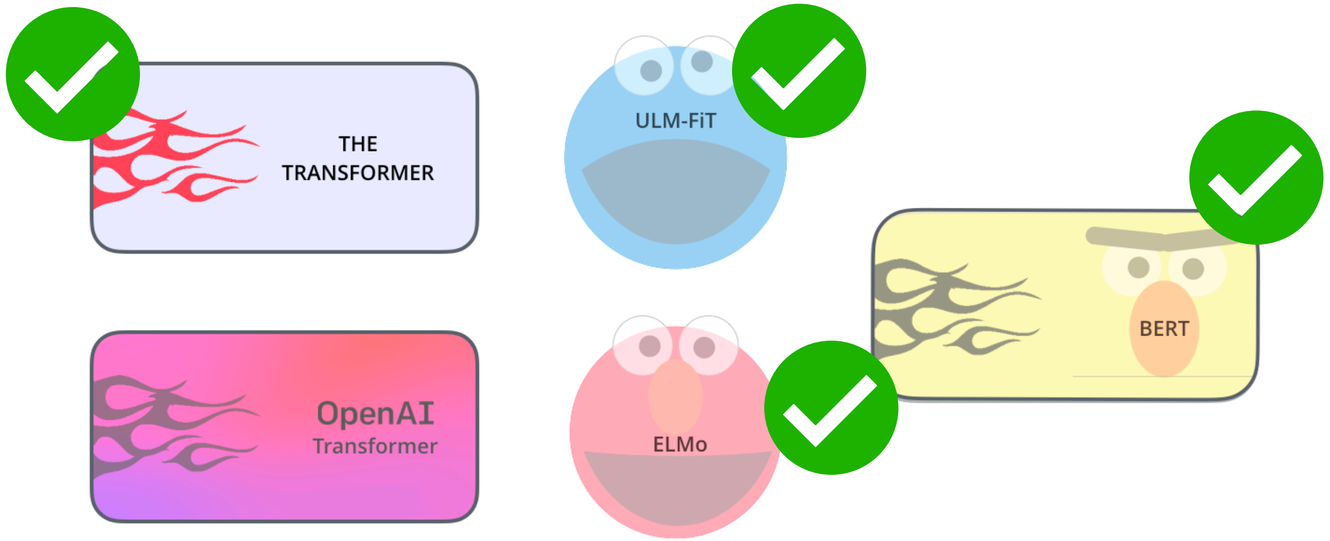
BERT

Pre-training of Deep
Bidirectional Transformers
for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT family... (RoBERT, RoBERTA, alBERT, ... Hasta un tal BETO existe)



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

Open AI Transformer

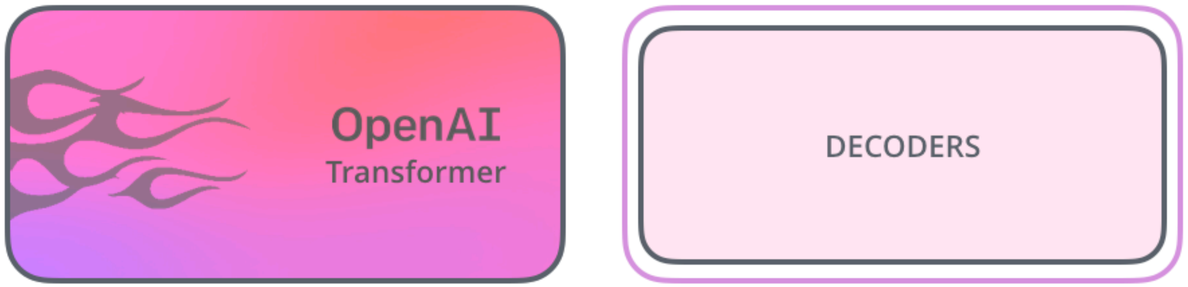
Pre-training a transformer Decoder for Language Modeling

Open AI

Se conforma con el codificador del transformer

Función de perdida:

- *NSP*



Open AI Transformer

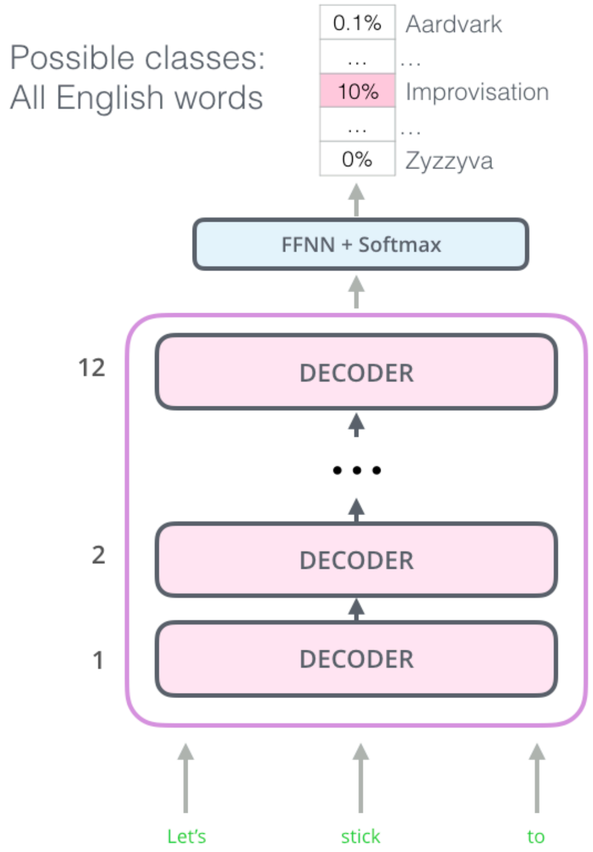
Pre-training a transformer Decoder for Language Modeling

Open AI

Se conforma con el codificador del transformer

Función de perdida:

- *NSP*



12 capas de decodificador

No tiene subcapas 'Encoder-decoder attention'

Tiene una capa de autoatención enmascarada

Entrenado en conjunto de datos de 7.000 libros

Open AI Transformer

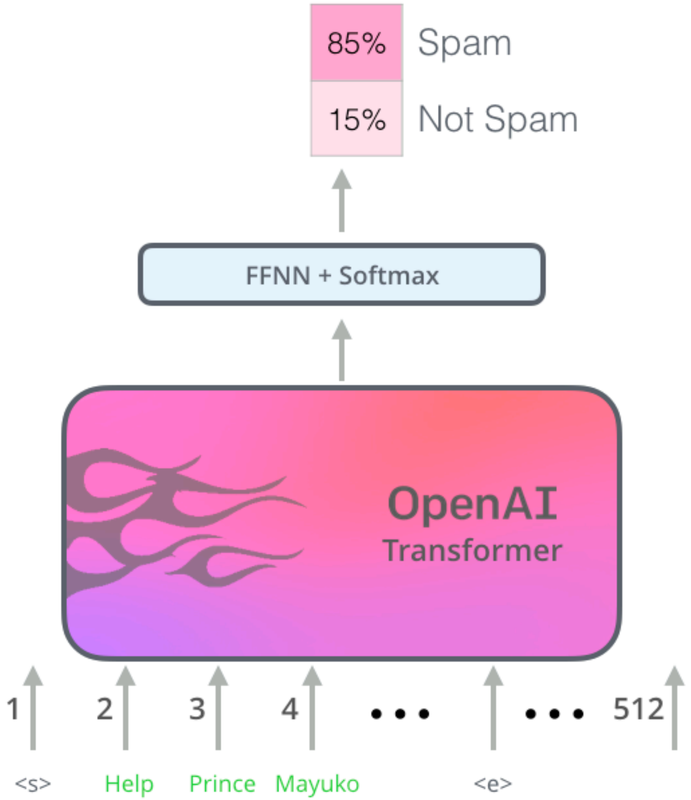
Pre-training a transformer Decoder for Language Modeling

Open AI

Se conforma con el codificador del transformer

Función de perdida:

- *NSP*



Aprendizaje por transferencia

Clasificación
Ej.: “Smap” o “No spam”

Open AI Transformer

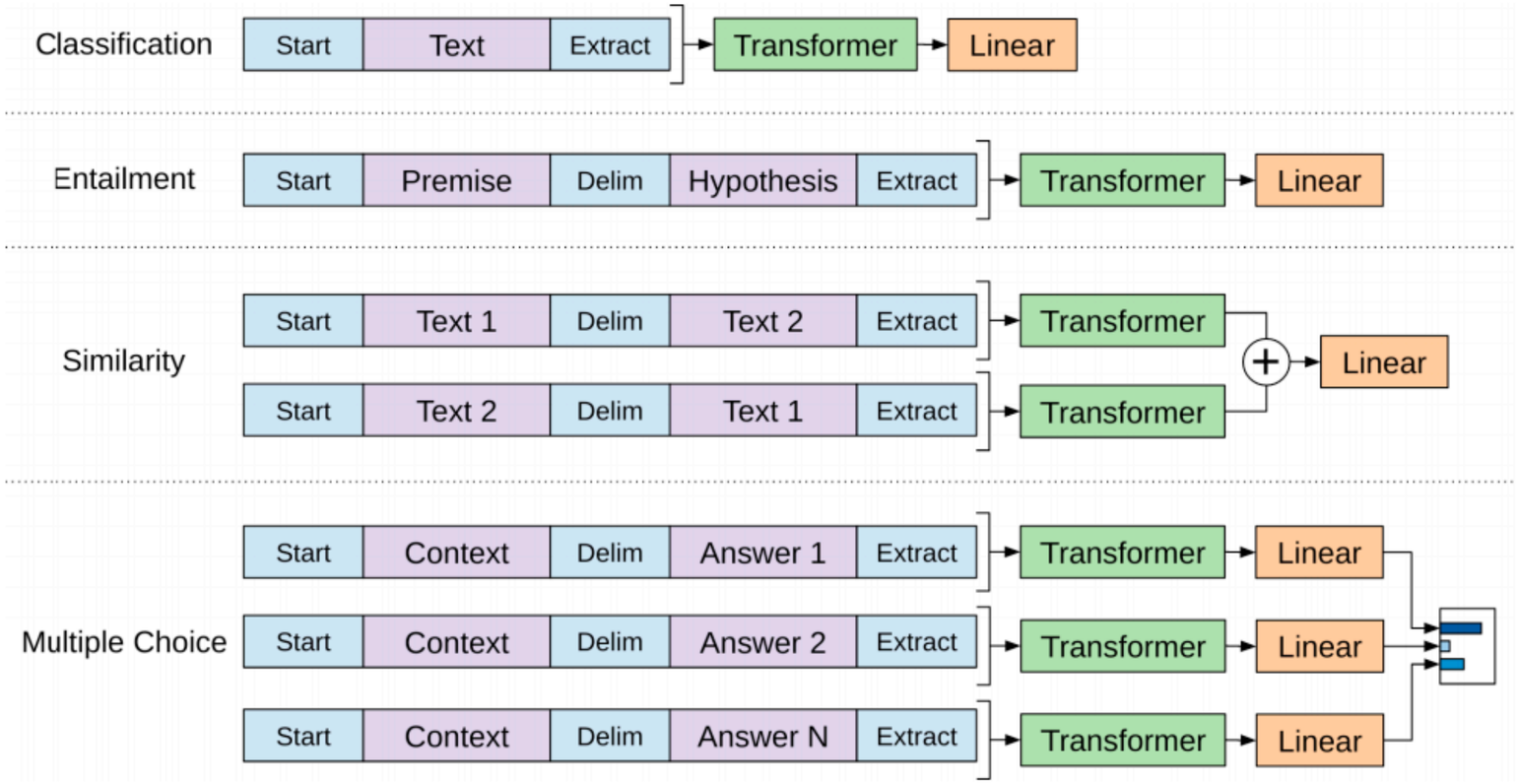
Pre-training a transformer Decoder for Language Modeling

Open AI

Se conforma con el codificador del transformer

Función de perdida:

- *NSP*



Transformaciones de entrada para manejar diferentes tareas

BERT

Bidirectional Encoder Representation for Transformers

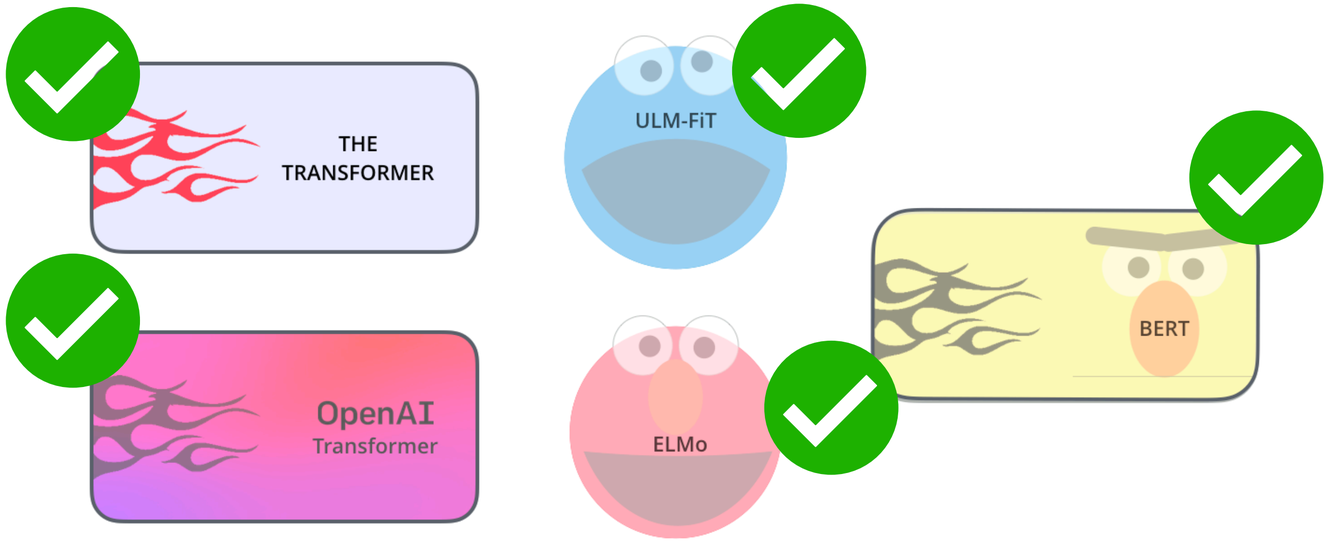
BERT

Pre-training of Deep
Bidirectional Transformers
for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT family... (RoBERT, RoBERTA, alBERT, ... Hasta un tal BETO existe)



(ULMFiT no tiene nada que ver con Cookie Monster. Pero no se me ocurre nada más...)

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

From Decoder to Encoders

¿Bidireccional? → *Enmascaramiento*

(Encoder)

LSTM → *Transformers*

Bidireccional

(Decoder)

LSTM → *Transformers*

Lenguaje directo

BERT

Bidirectional Encoder Representation for Transformers

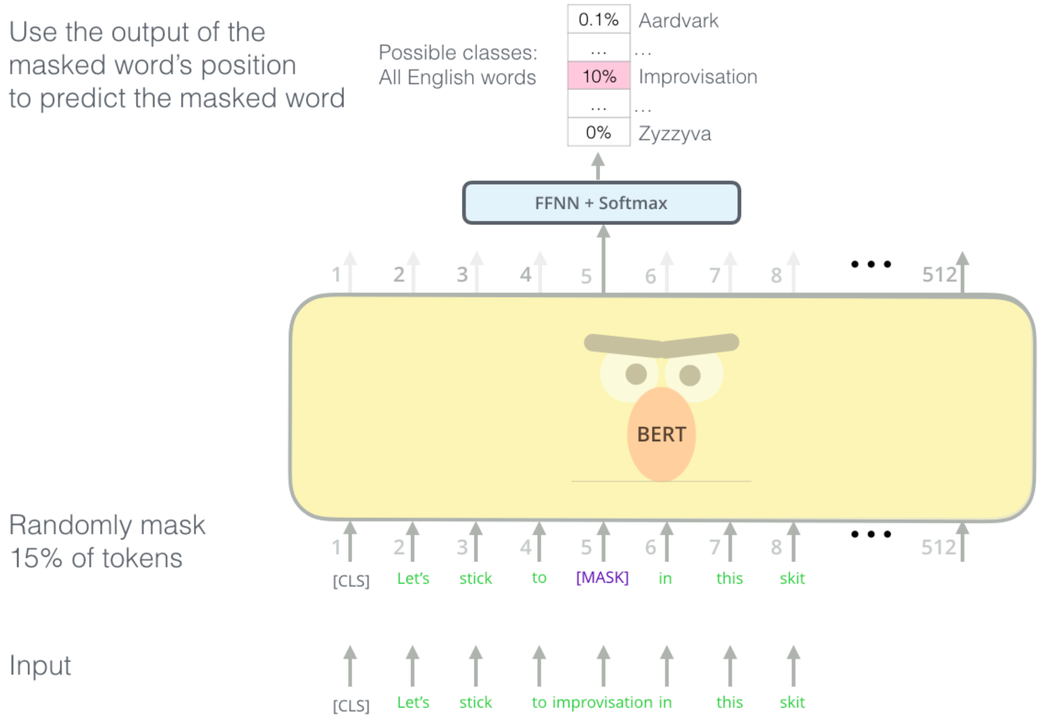
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Marked Language Model (modelo de lenguaje enmascarado)



BERT enmascara al 15% de palabras de entrada

Se le pide al modelo predecir la siguiente palabra

(Reemplazo aleatorio)

BERT

Bidirectional Encoder Representation for Transformers

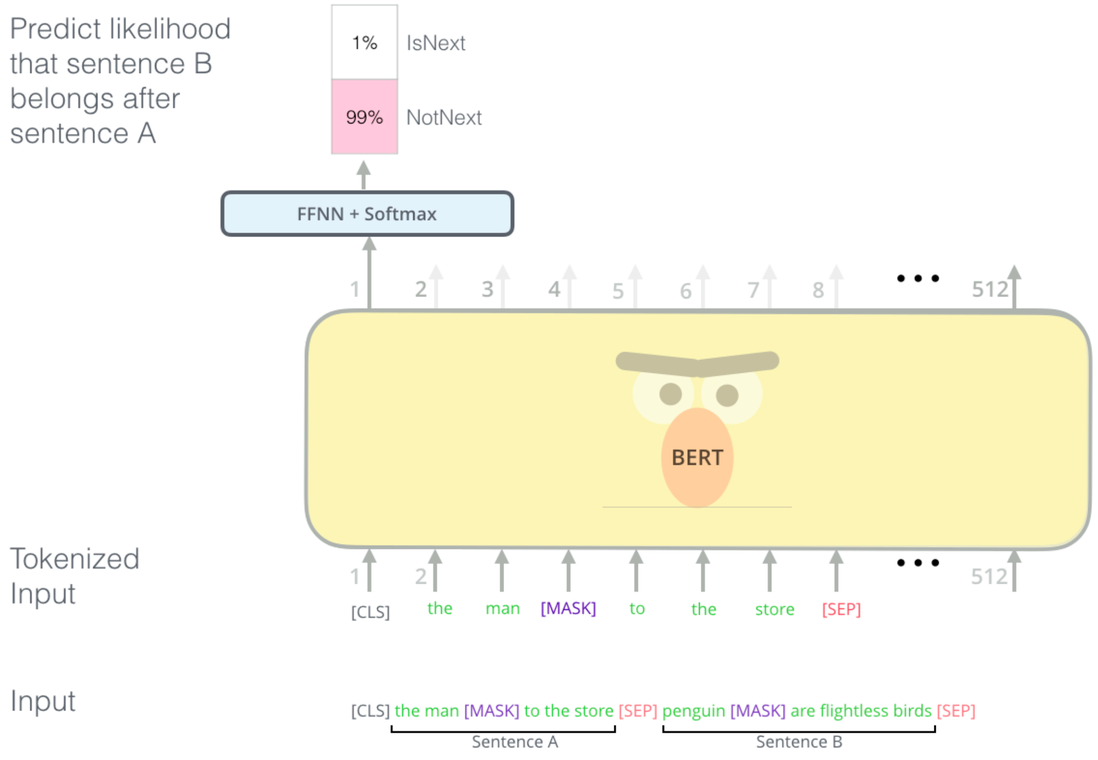
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Two-sentence Tasks (Tarea de dos oraciones)



Clasifica la siguiente oración. Si continua o no.

Bert usa WordPieces, no token por palabras.

BERT

Bidirectional Encoder Representation for Transformers

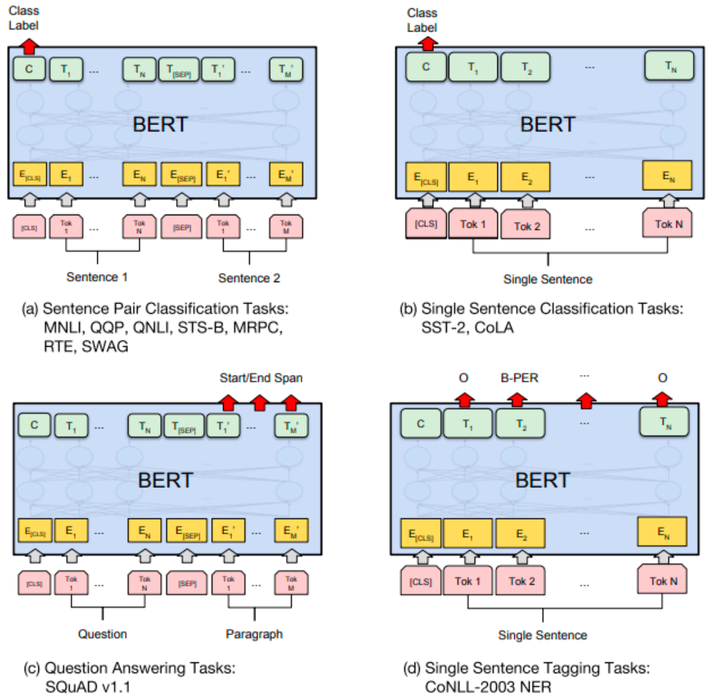
BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

Task specific-Models (BERT for different tasks)



Sentence pair classification task

Single sentence classification tasks

Question Answering tasks

Single sentence tagging tasks

BERT

Bidirectional Encoder Representation for Transformers

BERT

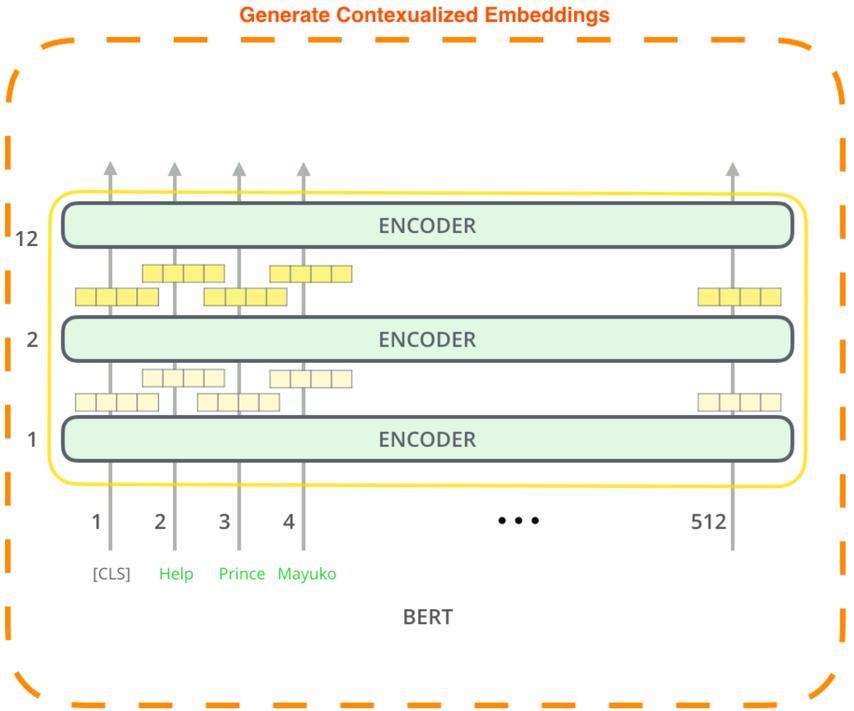
Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

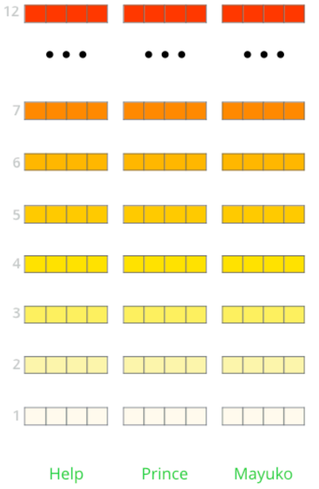
- *Mask ML*
- *NSP*

BERT for feature extraction

Like ELMo, with the pre-trained BERT create contextualized word embeddings



The output of each encoder layer along each token's path can be used as a feature representing that token.



But which one should we use?

BERT

Bidirectional Encoder Representation for Transformers

BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

Función de perdida:

- *Mask ML*
- *NSP*

BERT for feature extraction

Like ELMo, with the pre-trained BERT create contextualized word embeddings