

Deep Learning for Natural Language Processing

E. Wilderd Mamani Condori

Mcs student

errol.mamani@ucsp.edu.pe



Outline

- una vista a NLP
- una simple deep Network para classification (texto)
 - embeddings (word2vect)
 - convolution (CNN)
- RNNs para NLP
 - datos de secuencia(sequence data)
 - LSTM & GRU
- Arquitectura Transformers
 - BERT
- Aplicaciones

Natural Language Processing

Es el campo de estudio que trata de entender **cómo : funciona el lenguaje, se construye, como poder generar nuevo lenguaje** y todas las tareas relacionadas a esta.

Natural Language Processing

- text generation
- machine translation
- question answering
- speech recognition, entre otros



¿Que es texto?

tu tienes que pensar en texto como una secuencia de:

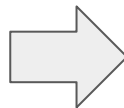
- caracteres
- **palabras (tokens)**
- frases y nombre de entidades
- oraciones
- párrafos
- ...

bag of words (sparse)

BOW representa: la suma de vectores esparsos (on-hot-encoding)

~100k columns

	good	movie	very	a	did	like
very →	0	0	1	0	0	0
good →	1	0	0	0	0	0
movie →	0	1	0	0	0	0

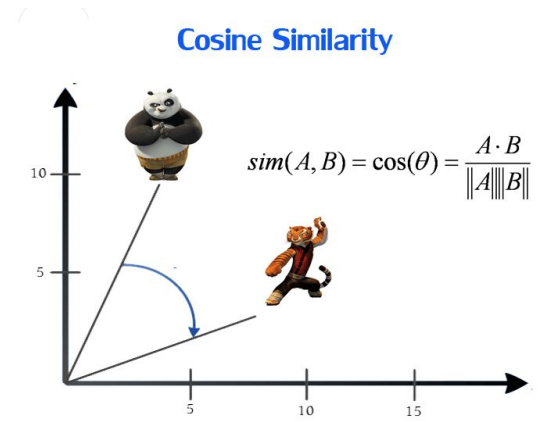
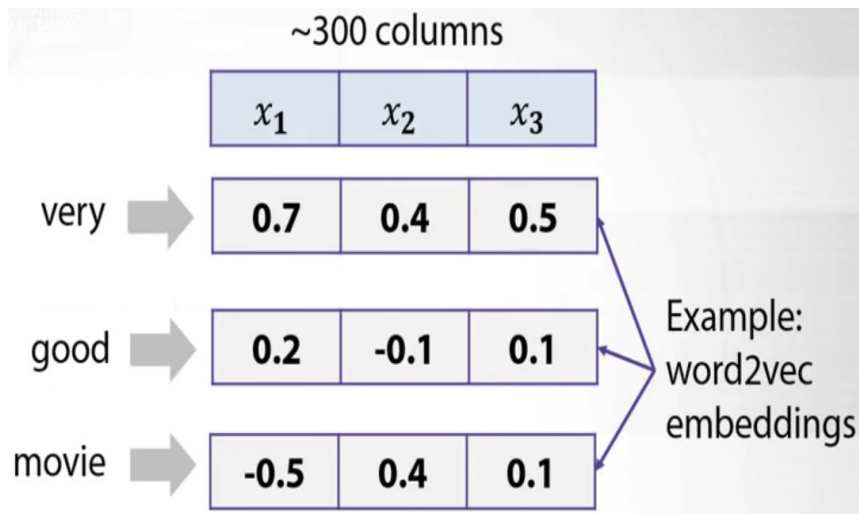


~100k columns

	good	movie	very	a	did	like
very →	0	0	1	0	0	0
			+			
good →	1	0	0	0	0	0
			+			
movie →	0	1	0	0	0	0
			=			
very good movie →	1	1	1	0	0	0

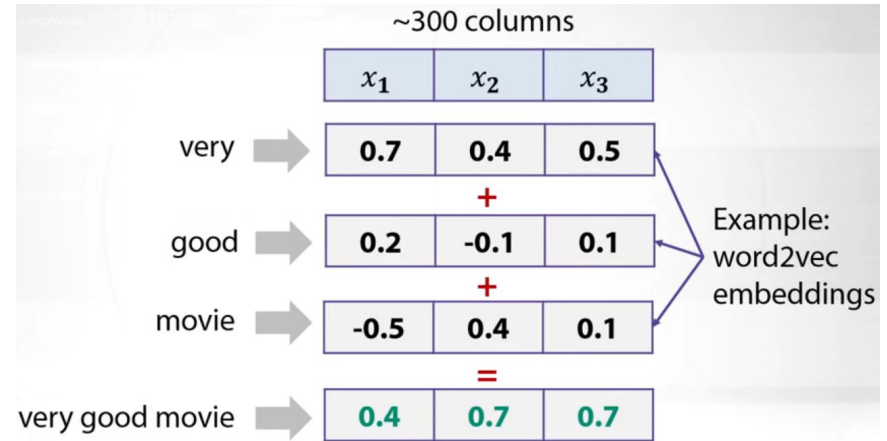
forma Neural (representación densa)

word2vec: “palabras que tienen similar contexto, tienden a tener vectores colineales” ([Tomas Mikolov in 2013 at Google.](#))

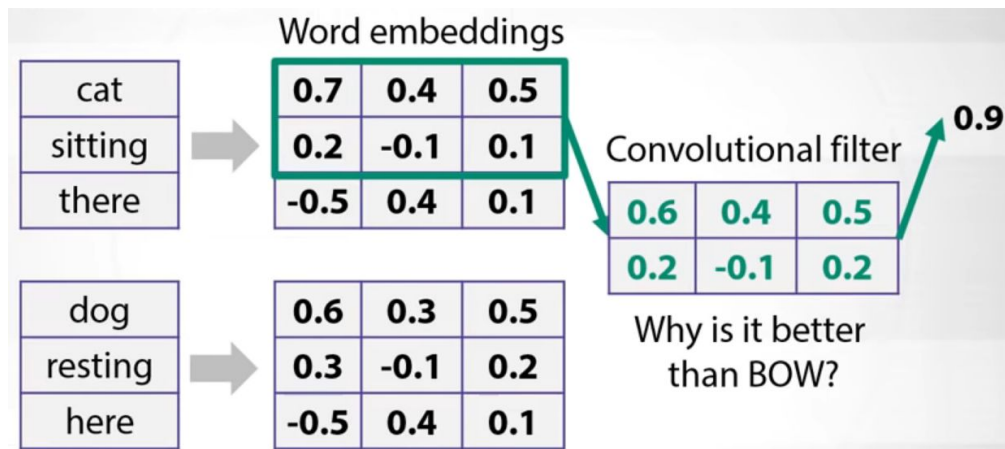


forma Neural (representación densa)

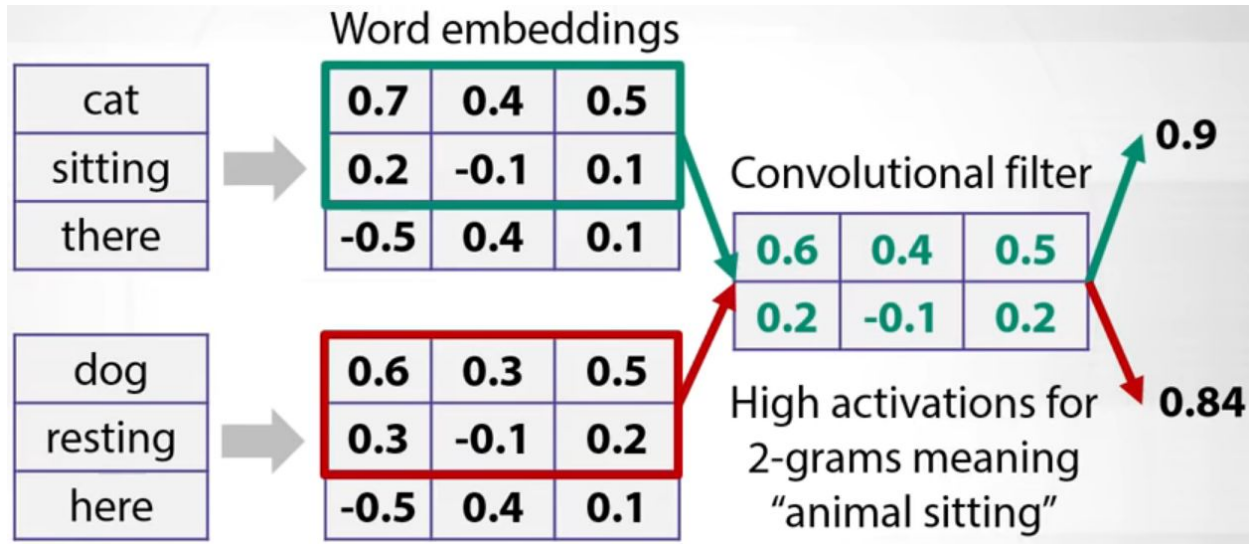
la suma de **word2vect** es ya un buen descriptor de texto



una mejor forma: 1D convolution



la convolution provee alta activación para 2-grams, “posee cierto significado”

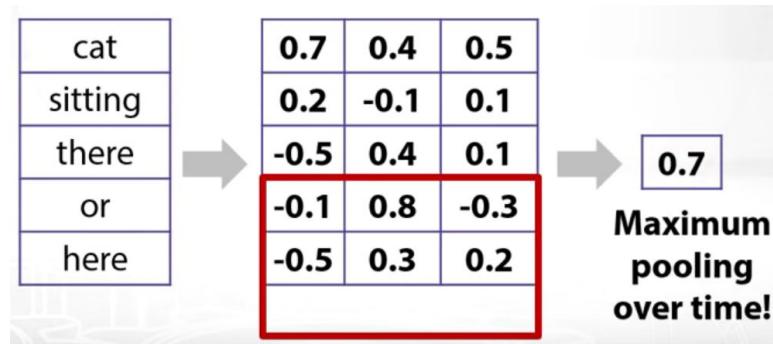
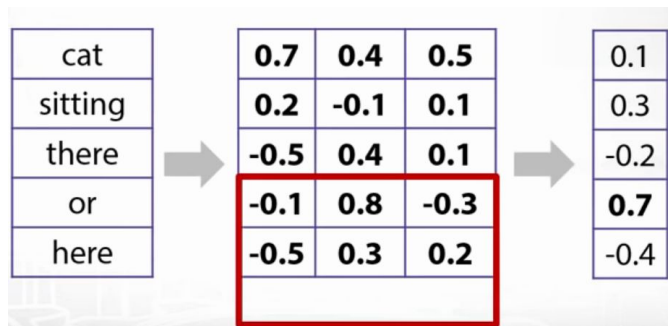


word2vect: similares palabras son similares en terminos de distancia cosenos(similar al producto punto)

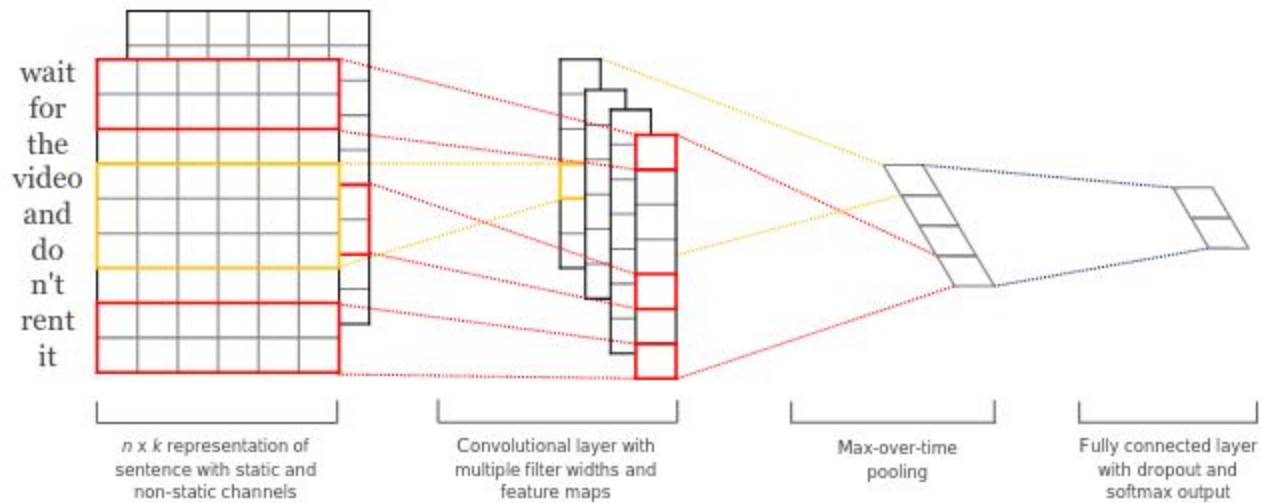
vamos hacer muchos filtros

el filtro se generaliza para n-grams(3,4,5,...,n)

1D: por que solo se desliza hacia abajo



CNN



Model architecture with two channels for an example sentence. CNN for sentences classification

yoon kim : NYK university 2014 paper : <https://arxiv.org/pdf/1408.5882.pdf>

RNNs (Redes Neuronales Recurrentes)

- red neuronal ,bueno modelando **datos secuenciales**.



¿podemos predecir hacia qué lugar
se mueve?

RNNs (Redes Neuronales Recurrentes)

- **datos secuenciales.**



**¿podemos predecir hacia qué
lugar se mueve?**

¿como podria saber?

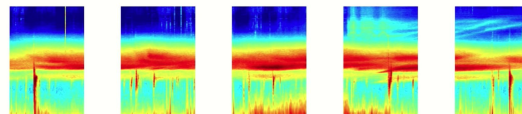


¿no lo entiendo!!?

Datos secuenciales



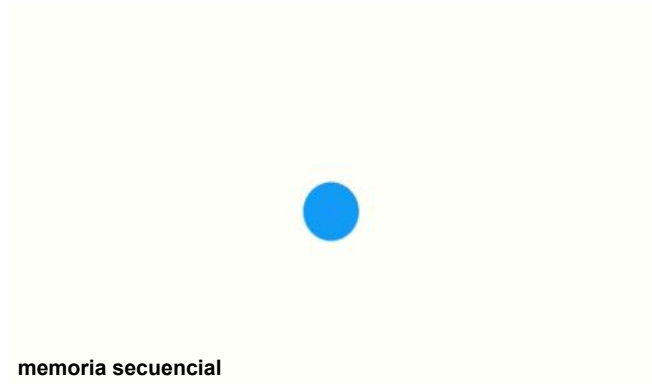
esto viene en
muchas formas



Espectrogramas (audio en trozos)

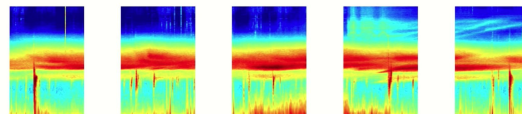
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Datos secuenciales



esto viene en
muchas formas

La memoria secuencial es un mecanismo
que facilita que su cerebro reconozca los
patrones de secuencia



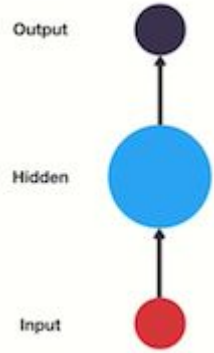
Espectrogramas (audio en trozos)

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

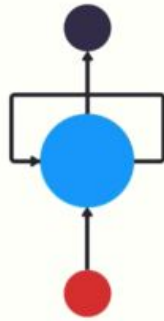
Z Y X W V U T S R Q P O N M L K J I H G F E D C B A

E...?

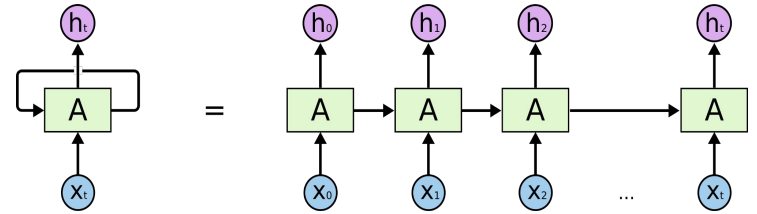
usar información previa para efectuar las
posteriores



feed forward Neural Network



RNN:



RNN en varios instantes de tiempo, Image Credit— colah's blog

ejemplos en NLP



chatbot: clasifican intención de los usuarios

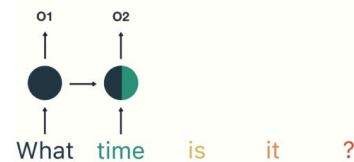
What time is it?



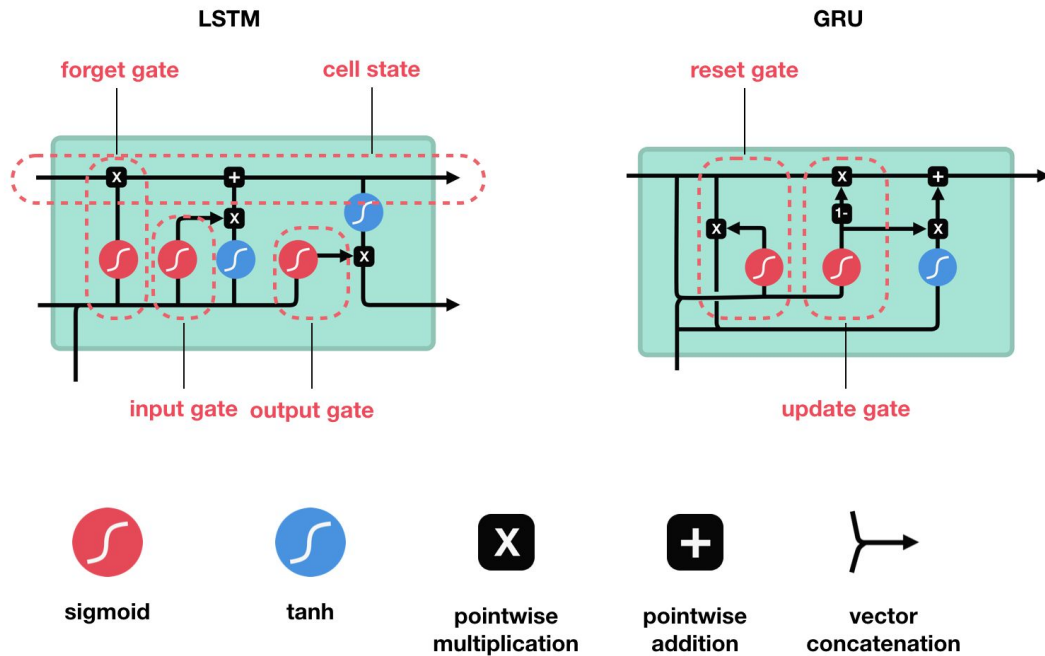
**vanishing
gradient**



estado final hidden de un
RNN



LSTM & GRU



LSTM & GRU

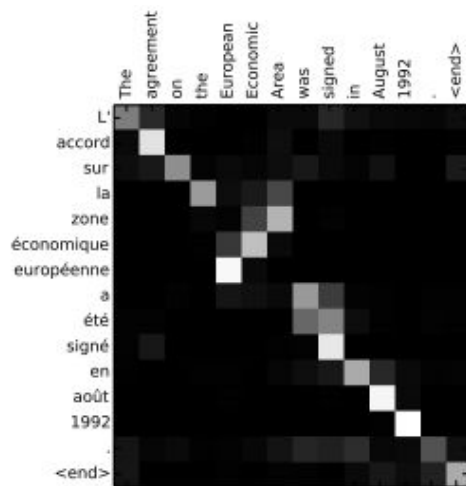
- GRU reduce el número de operaciones tensoriales comparado con LSTM.
- computacionalmente rápido y menos costoso que LSTM
- la principal diferencia son los gates [Cho, et al. \(2014\)](#)

Transformers

por lo que necesitamos ?

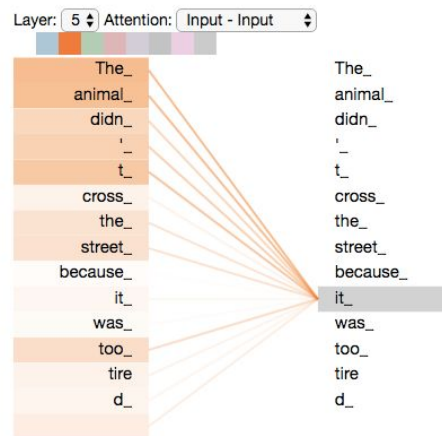
- Los modelos de Gate como LSTM o GRU son para propagación de errores de largo alcance.
- La secuencialidad prohíbe la paralelización dentro de las instancias.
- Las dependencias de largo alcance siguen siendo complicadas, a pesar de las puertas.
- Los estados alineados con la secuencia en RNN son un desperdicio, son difíciles de modelar, etc.
- RNN / CNN manejan secuencias palabra por palabra secuencialmente, lo cual es un obstáculo para el paralelismo

Mecanismo de atención



Matriz de doble con 2 oraciones columna y fila

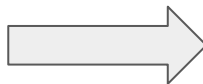
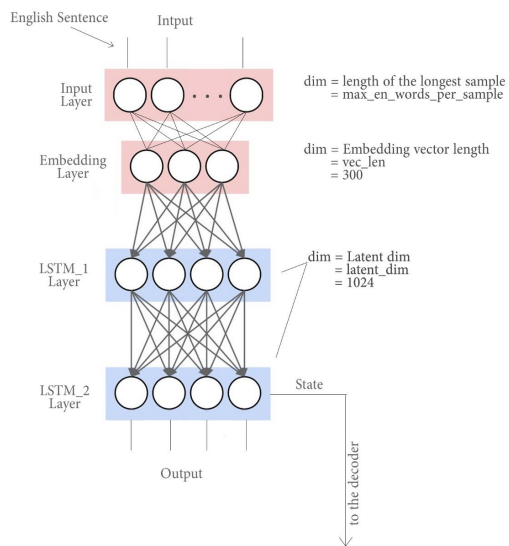
"The animal didn't cross the street
because it was too tired"



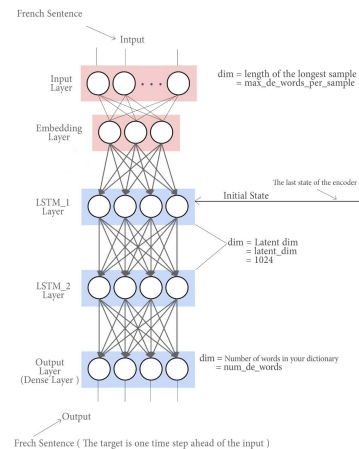
mecanismo de self attention

seq2seq

The Encoder



The Decoder





seq2seq model

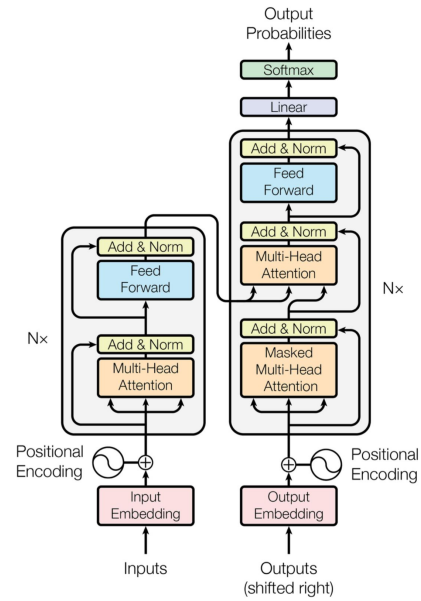
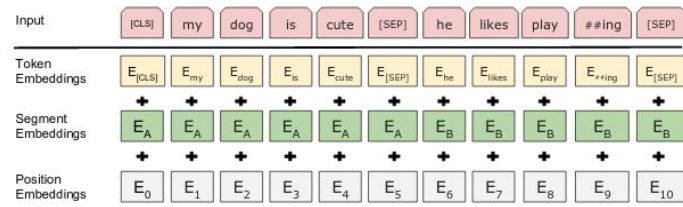


Figure 1: The Transformer - model architecture.

BERT



bert: representación de input

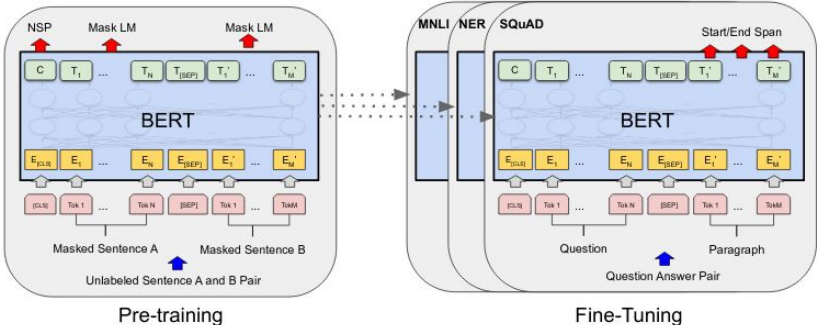


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

Algunas Aplicaciones

- <https://talktotransformer.com/>
- <https://gpt2.apps.allenai.org/>



GPT-2 Explorer

This demonstration uses the public **345M** ^{117M} parameter **OpenAI GPT-2** language model to generate sentences.

Enter some initial text and the model will generate the most likely next words. You can click on one of those words to choose it and continue or just keep typing. Click the left arrow at the bottom to undo your last choice.

Sentence:

Joel is

Options:

18.0% **a**
8.4% **the**
5.1% **an**
2.8% **also**
2.4% **one**
2.2% **not**
1.6% **currently**
1.3% **joined**
1.3% **on**
1.1% **now**
-- Undo

Built at the **Allen Institute for Artificial Intelligence** using Hugging Face's **pytorch-pretrained-BERT** library. See the [source code on GitHub](#).

