Transformers: BERT

Orlando Ramos Flores

Contenido

- BERT
  - ¿Qué es BERT?
  - Motivación
  - Masked Language Model

BERT

¿Qué es BERT?

- BERT (Bidirectional Encoder Representations from Transformers) es nuevo modelo de representación del lenguaje, y está diseñado para pre-entrenar representaciones bidireccionales profundas a partir de texto sin etiquetar mediante el condicionamiento conjunto del contexto izquierdo y derecho en todas las capas.
- Como resultado, el modelo BERT pre-entrenado se puede ajustar con solo una capa de salida adicional para crear modelos de última generación para una amplia gama de tareas.

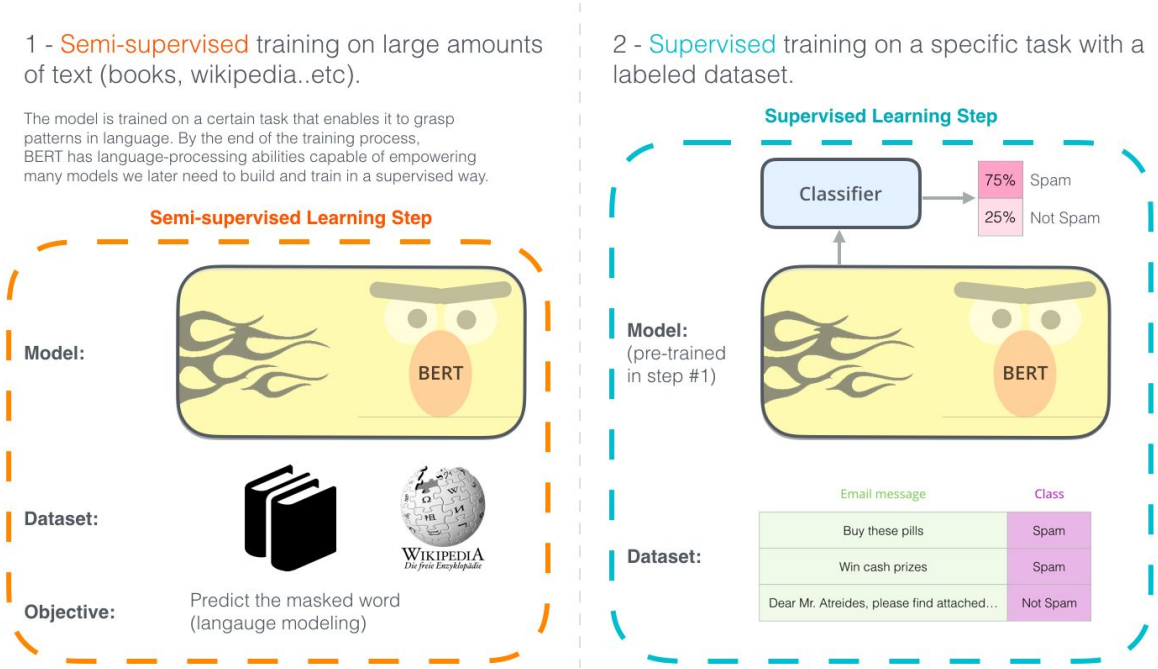
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT: Base y Large

- Sea L el número de capas (bloques de Transformers). El tamaño del estado oculto denotado por H, y el número de cabezas de self-attention como A.
- $BERT_{BASE}$
  - L = 12
  - H = 768
  - A = 12
  - Parámetros = 110M
- $BERT_{LARGE}$
  - L = 24
  - H = 1024
  - A = 16
  - Parámetros = 340M

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT

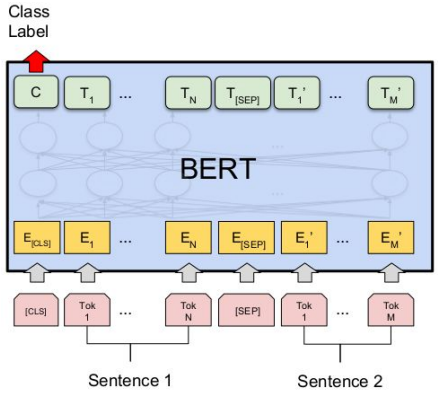


Los dos pasos de cómo se desarrolla BERT. Se puede descargar el modelo pre-entrenado en el paso 1 (entrenado con datos no anotados) y solo preocuparse por ajustarlo para el paso 2.

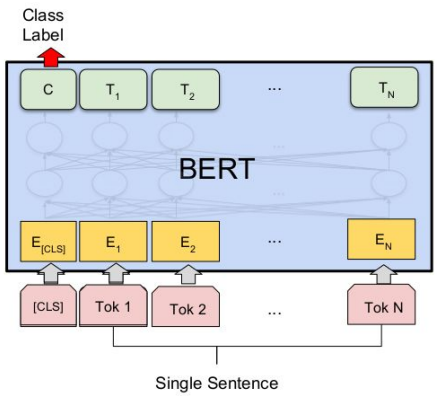
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Fine-tuning BERT

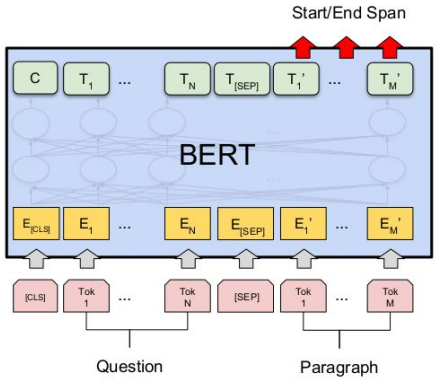
- Obtiene nuevos resultados de última generación en once tareas de procesamiento de lenguaje natural,
- Incluido el aumento de la puntuación GLUE al 80,5 % (7,7 % de mejora absoluta),
- La precisión de MultiNLI al 86,7 % (4,6 % de mejora absoluta),
- SQuAD v1.1 Question & Answering Test F1 a 93.2 (1.5 puntos de mejora absoluta) y
- SQuAD v2.0 Test F1 a 83.1 (5.1 puntos de mejora absoluta).



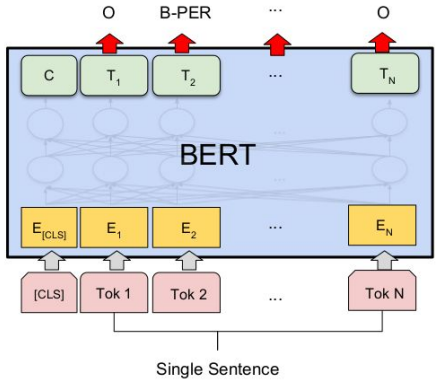
(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG



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



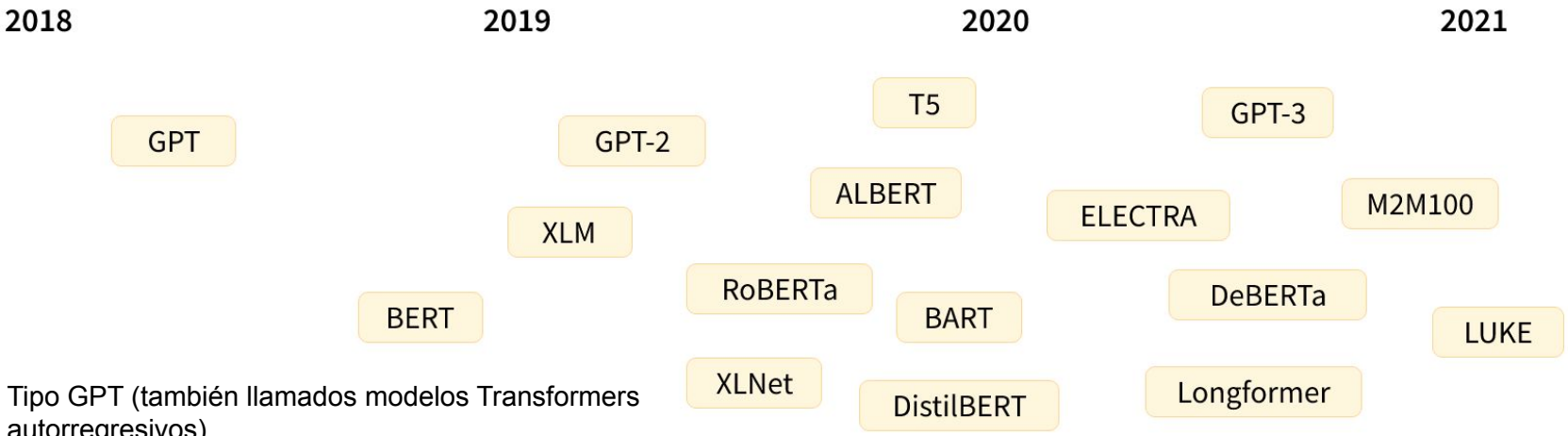
(c) Question Answering Tasks: SQuAD v1.1



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

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Modelos de Transformers: Historia



- Tipo GPT (también llamados modelos Transformers autorregresivos).
- Tipo BERT (también llamados modelos Transformers de codificación automática).
- Tipo BART/T5 (también llamados modelos de Transformers de secuencia a secuencia)

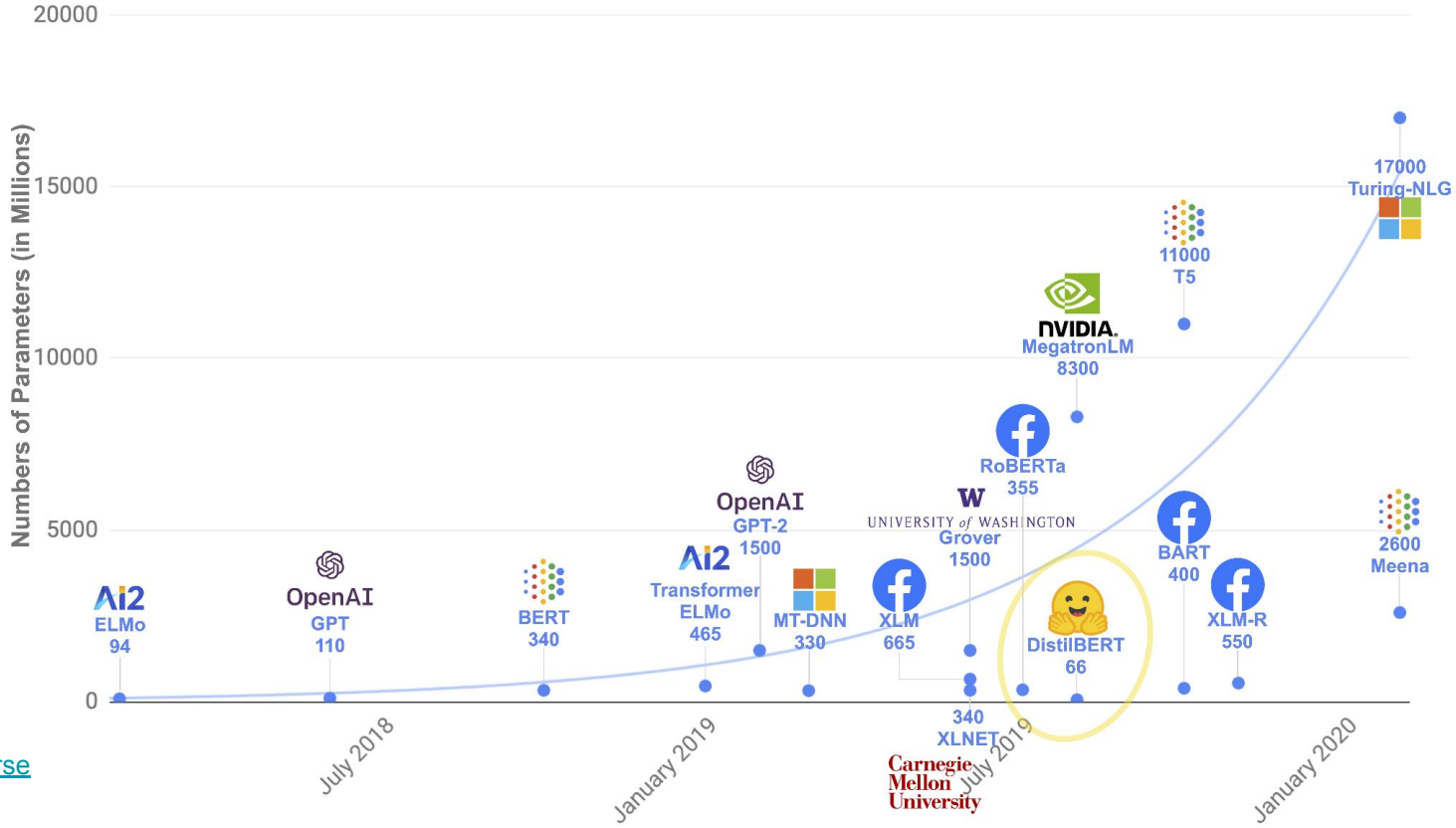
Hugging Face Course

Modelos de Transformers: Historia

- Todos los modelos de Transformer mencionados anteriormente (GPT, BERT, BART, T5, etc.) han sido entrenados como **modelos de lenguaje**.
- Esto significa que han sido entrenados sobre grandes cantidades de texto sin procesar de manera auto supervisada.
- El aprendizaje autosupervisado es un tipo de entrenamiento en el que el objetivo se calcula automáticamente a partir de las entradas del modelo.
- Significa que no se necesitan humanos para etiquetar los datos

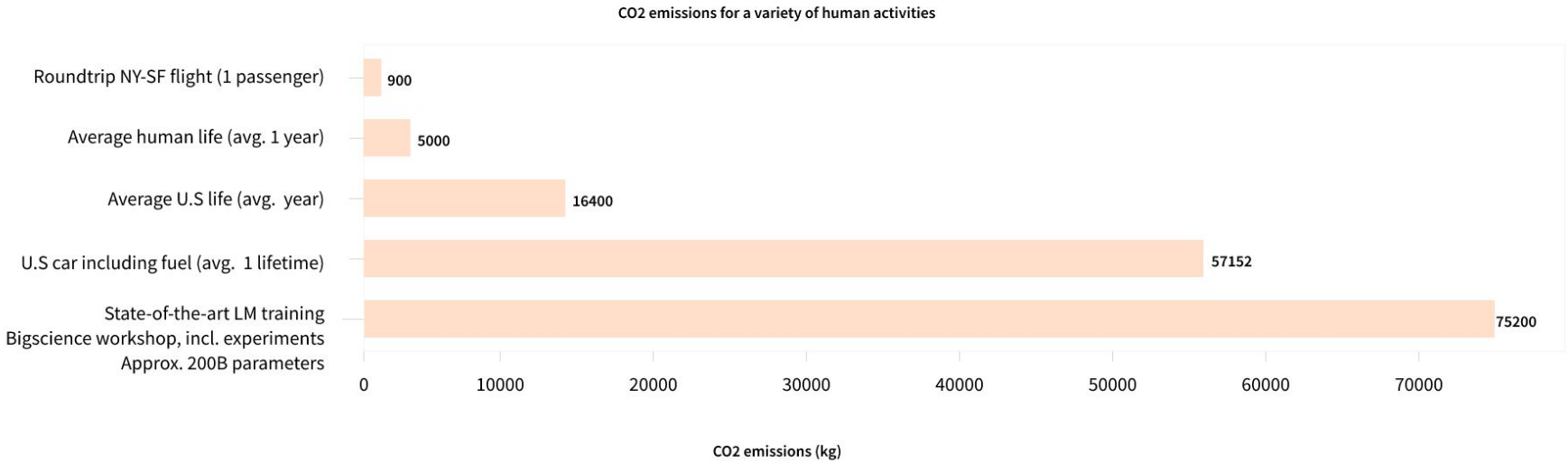
Hugging Face Course

Transformers



[Hugging Face Course]

Transformers: Huella de carbono



Hugging Face Course
The carbon footprint of Transformers

Motivación

BERT: Motivación

- Hay dos estrategias existentes para aplicar **representaciones de lenguaje previamente entrenadas** a *downstream tasks*[1]: featured-based (basadas en características) y fine-tuning (de ajuste fino).
- El enfoque featured-based con ELMo utiliza arquitecturas específicas de tareas que incluyen las representaciones pre-entrenadas como características adicionales.
- El enfoque fine-tuning tal como Generative Pre-trained Transformer (OpenAI GPT) presenta parámetros mínimos específicos de la tarea y se entrena en *downstream tasks*[1] simplemente ajustando todos los parámetros pre-entrenados.

1. Downstream tasks: en el campo del PLN se llaman a aquellas tareas de aprendizaje supervisado que utilizan un modelo o componente pre-entrenado

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT: Motivación

- Los dos enfoques comparten la misma función objetivo durante el pre-entrenamiento, donde usan **modelos de lenguaje unidireccionales** para aprender representaciones generales del lenguaje.
- En el paper BERT argumentan que tales técnicas restringen el poder de las representaciones pre-entrenadas, especialmente para los enfoques de fine-tuning.
- La principal limitación es que los modelos de lenguaje estándar son unidireccionales, y esto limita la elección de arquitecturas que se pueden usar durante el pre-entrenamiento.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BERT: Motivación

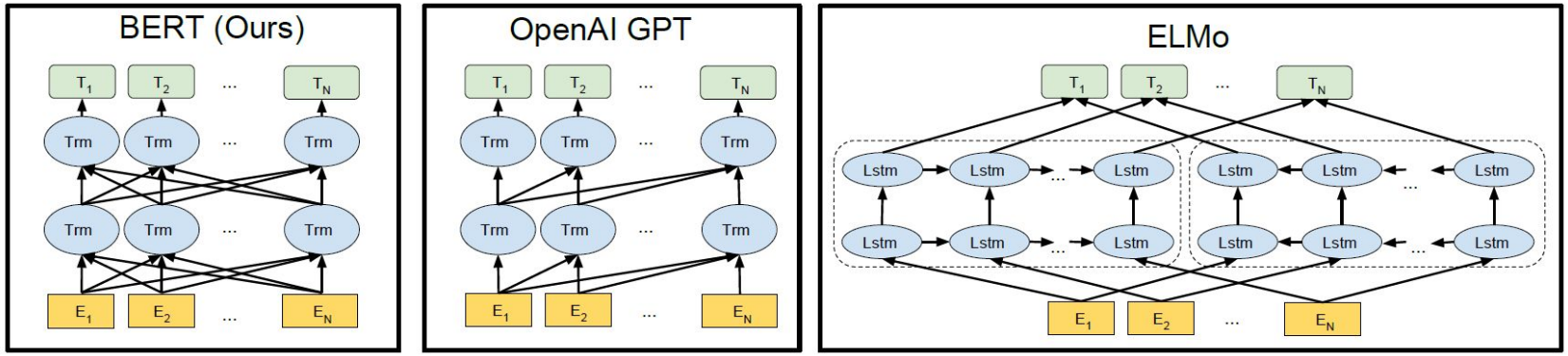


Figure 3: Differences in pre-training model architectures. BERT uses a bidirectional Transformer. OpenAI GPT uses a left-to-right Transformer. ELMo uses the concatenation of independently trained left-to-right and right-to-left LSTMs to generate features for downstream tasks. Among the three, only BERT representations are jointly conditioned on both left and right context in all layers. In addition to the architecture differences, BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Masked Language Model

Masked Language Model (MLM)

- Intuitivamente, es razonable creer que un modelo bidireccional profundo es estrictamente más poderoso que un modelo de izquierda-a-derecha o que uno con concatenación superficial de un modelo de izquierda-a-derecha y de derecha-a-izquierda.
- Desafortunadamente, los modelos de lenguaje condicional estándar solo se pueden entrenar de izquierda-a-derecha o de derecha-a-izquierda, ya que el condicionamiento bidireccional permitiría que cada palabra se “viera a sí misma” indirectamente, y el modelo podría predecir trivialmente la palabra objetivo en un formato de contexto de varias capas.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Masked Language Model (MLM)

- Entonces, para entrenar una representación bidireccional profunda, simplemente se enmascara algún porcentaje de los tokens de entrada al azar y luego se predicen esos tokens enmascarados.
- Este procedimiento se refiere como un model “masked LM” (MLM), aunque a menudo se lo denomina tarea *Cloze* en la literatura.
- En este caso, los vectores ocultos finales correspondientes a los tokens de máscara se introducen en un softmax de salida sobre el vocabulario, como en un LM estándar.
- En todos los experimentos, enmascaran el 15 % de todas las piezas de palabras en cada secuencia al azar.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Taylor, W. L. (1953). “Cloze procedure”: A new tool for measuring readability. Journalism quarterly, 30(4), 415-433.