RESUMEN del ARTÍCULO BERT:

*Pre-training of Deep Bidirectional Transformers for Language Understanding*

## Introducción

El artículo presenta un nuevo modelo para el procesamiento del lenguaje natural (NLP) el cual obtiene resultados de vanguardia en 11 tareas diferentes.

Existen dos tipos de estrategias para la representación de modelos NLP preentrenados; las basadas en características y las basadas en fine tuning. BERT es un modelo preentrenado basado en fine tunning. Por lo tanto, **BERT se distingue en dos fases: fase de preentrenamiento y fase de “fine tuning”** o ajuste.

## Arquitectura del modelo

BERT tiene **una arquitectura de transformer bidireccional multi capa**. Los autores afirman que su modelo es casi idéntico a la implementación original de Vaswani et al. (2017).

Se diferencian dos implementaciones, el **BERT-base** y el **BERT-largo**, en los cuales se define la cantidad de capas o bloques de Transformers como L, el tamaño oculto como H y el número de cabezas de autoatención como A. BERT-base tiene L= 12, H= 768 y , A= 12 con un total de 110 M parámetros, y BERT-largo tiene L= 24, H= 1024, y A= 16 con un total de 340 millones de parámetros.

## Fase de preentrenamiento

La fase de **preentrenamiento**, la cual es **entrenada en datos no etiquetados**, ha demostrado ser útil en tareas como la inferencia del lenguaje natural, la predicción entre la relación de frases a nivel holístico, y tareas a nivel de token ( la respuesta de preguntas por ejemplo), sin embargo, en los modelos previos como CHAT GPT de open AI, en la fase de preentramiento los modelos eran entrenados unidireccionalmente de izquierda a derecha, es decir, durante la fase de inferencia y entrenamiento los modelos solo tienen acceso a lo que hay antes ( a la izquierda ) de la respuesta que se quiere predecir.

Lo que diferencia el modelo BERT del resto de transformes es que **en la etapa preentrenamiento usa un enfoque bidireccional**; el modelo tiene acceso a información tanto a antes como después de la palabra o respuesta que es objetivo de inferencia. Para esta fase de preentrenamiento se utiliza un MLM (masked languaje model) o modelo de lenguaje enmascarado, el cual enmascara unas palabras seleccionadas aleatoriamente del input del modelo y su objetivo es predecir las palabras originales del texto, basándose únicamente en el contexto. Además del MLM, el preentrenamiento también incluye la tarea de predicción de la siguiente frase.

Durante esta primera fase, el modelo es preentrenado en datasets de más de 800 y 2500 millones de palabras, ignorando tablas, títulos, listas, y encabezados, los autores hacen hincapié en que el uso de texto a nivel documento es crítico para la extracción de secuencias contiguas extensas.

## Fase de fine tuning

En esta segunda fase, el mecanismo de autoatención permite a BERT realizar tareas que involucran texto único o pares de texto intercambiando las entradas y salidas apropiadas. Para cada tarea, se conectan las entradas y salidas específicas de la tarea y se hace el fine tuneado de los parámetros.

**En la entrada**, las oraciones A y B del preentrenamiento son análogas a (1) pares de oraciones en la paráfrasis, (2) pares hipótesis-premisa en la implicación, (3) pares de pregunta-párrafo en la respuesta a preguntas, y (4) un par de texto generado en el etiquetado de secuencias. **En la salida**, las representaciones de tokens se alimentan a una capa de salida para tareas a nivel de token, como el etiquetado de secuencias o la respuesta a preguntas, y la clasificación se alimenta a una capa de salida, como la implicación o el análisis de sentimiento.

**Esta segunda fase es relativamente económica** permitiendo replicar los resultados en poco tiempo con los recursos adecuados.

## Experimentos y fase de ablación

Se ha experimentado con 4 tareas generales que involucran 11 diferentes subtareas, tales como, evaluación del entendimiento general del lenguaje, respuesta a preguntas, y la continuación de sentencias. En todas las tareas BERT-largo bate métricas preestablecidas. Para cada tarea el modelo a sido fine tuneado con diferentes tamaños de lote y número de épocas.

La **fase de ablación** remarca los efectos **y la importancia del preentrenamiento**, en la cual se le da especial importancia a la bidireccionalidad del modelo, **la importancia del tamaño del modelo**, que demuestra que por lo general los modelos más extensos obtienen mejores resultados en las métricas, incluso para datasets pequeños. Por último, el **enfoque basado en características** ofrece ciertas ventajas, como beneficios computacionales, o el hecho de no tener que requerir de una arquitectura específica para cada tarea.

## Conclusión

La **parte no supervisada** del preentrenamiento es una **parte fundamental** para la mayoría de las tareas del entendimiento del lenguaje, además, los autores afirman que su mayor contribución ha sido lograr la generalización de una arquitectura profunda bidireccional permitiendo aplicar el mismo modelo a diversas tareas NLP con éxito.

## Posible mejora

Desde mi desconocimiento en estos temas veo difícil de aplicar una mejora en este tipo de modelos, sin embargo, se me ocurre una posible mejora; dada la cantidad de parámetros que maneja el modelo (340M), tal vez, **una fase de preprocesado en la que se realiza una reducción de dimensionalidad** podría eliminar palabras redundantes que no aportan mucha información en el entendimiento del texto. De esta manera se podría eliminar ruido del modelo y podría generalizar mejor además de aumentar la rapidez del entrenamiento, aunque esto conlleve cierta pérdida de información.