Javier Soto

Reconocimiento Visual con Deep Learning
Quiz #3

1.-

Diferencias:

Uso principal

Los modelos de atención se utilizan principalmente en tareas de procesamiento de lenguaje, como traducción automática, generación de texto y resumen de texto. En cambio, los modelos convolucionales se utilizan principalmente en tareas de visión por computadora, ya que son buenos para detectar patrones locales en imágenes.

Datos de Entrada y Salida

La entrada a un modelo de atención es generalmente una secuencia de datos, como una serie de palabras, y la salida puede variar según la tarea, pero generalmente involucra generar una secuencia de salida, como una traducción. Por otro lado, la entrada a un modelo convolucional suele ser una imagen (matriz de píxeles), y la salida tiende a ser una tarea, como por ejemplo en la tarea 1, la salida es una clasificación de objetos en una imagen.

2.-

El las queries (Q) son representaciones de la entrada que se utilizarán para calcular la atención. Se obtienen aplicando una transformación lineal a la entrada original multiplicando la entrada por una matriz de pesos Q, que es un parámetro entrenable del modelo. Las keys (K) se obtienen de esta misma manera, pero se utilizan para calcular la relevancia entre las queries y las keys. Al igual que las queries y las keys, los values (V) se obtienen aplicando una transformación lineal a la entrada original utilizando una matriz de pesos. Los values son las representaciones de los datos de entrada que se utilizarán para calcular el resultado de la atención.

3.-

En el **self-attention**, un modelo se enfoca en calcular la atención dentro de una única secuencia de entrada, generalmente una frase o un documento. El objetivo es capturar las relaciones entre las palabras o tokens dentro de la misma secuencia. Cada token en la secuencia actúa tanto como una query como una key. Es decir, se utiliza para calcular la atención con respecto a todos los demás tokens en la misma secuencia. Los values son las representaciones de los tokens de entrada y se utilizan para calcular la salida de atención ponderada.

Ahora en el **cross-attention**, se calcula la atención entre dos secuencias diferentes. Esto es común en tareas de traducción automática, donde una secuencia es la frase en el idioma fuente y la otra es la frase en el idioma destino. Lo que hace es alinear las palabras o tokens en ambas secuencias. Las queries vienen de la secuencia de destino y se utilizan para calcular la atención con la secuencia de origen. Las keys provienen de la secuencia de origen y se utilizan para calcular la atención en relación con la secuencia de destino. Y los values son las representaciones de la secuencia de origen.

4.-

La principal diferencia entre aplicar un modelo de atención tipo Transformer a texto y a imágenes radica en la naturaleza de los datos de entrada y cómo se manejan. En el procesamiento de texto, los datos de entrada son secuencias de palabras o tokens. Cada palabra o token tiene una posición secuencial y relaciones específicas con otras palabras dentro de la secuencia. En cambio, en el procesamiento de imágenes, los datos de entrada son matrices de píxeles que representan la información visual. A diferencia del texto, no existe una secuencia estructural natural en las imágenes. Las relaciones entre píxeles no son tan directas como las relaciones entre palabras en una secuencia de texto.

Para resolver esta diferencia, en el procesamiento de texto se aplica el Transformer directamente y se utiliza en tareas de procesamiento de lenguaje natural (traducción, generación de texto, etc). Para procesar imágenes con un Transformer, es necesario convertir la información visual en una forma que el modelo pueda entender. Esto se logra a través de arquitecturas como Vision Transformers o Convolutional Transformers.

5.-

El concepto de Positional-Encoding es fundamental en los modelos de atención, como los Transformers, para que el modelo pueda manejar la información de la

secuencia en el orden correcto. Ya que los modelos de atención no tienen la capacidad interna de tener el orden correcto de los elementos en una secuencia, se utiliza el Positional-Encoding para proporcionar al modelo información sobre la posición relativa de cada elemento en la secuencia.

6.-

El objetivo principal de aplicar la operación softmax sobre el resultado QK^T (producto punto entre las queries y las keys transpuestas) en los modelos de atención es convertir estos empeoramientos en una distribución de probabilidad. Lo importante es que esta distribución de probabilidad refleje la importancia relativa de los elementos de la entrada con respecto a una query especifica. Esto permite al modelo dirigir su atención de manera efectiva hacia la información más relevante para la tarea que está realizando.