Universidad del Valle de Guatemala
Dep. Ciencias de La Computación
Deep Learning
01/09/2024

Gabriel García - 21352
Luis Pedro Montenegro - 21699

Hoja de Trabajo - 2

Ejercicio 2

1. **¿Cuál es la principal innovación de la arquitectura Transformer?**

   La innovación es que la arquitectura Transformer depende solamente de atención, prescindiendo del uso de convoluciones y recurrencia, para marcar dependencias globales entre sus entradas y salidas.

2. **¿Cómo funciona el mecanismo de atención de scaled dot-product?**

   Al tener una entrada de valores de dimensión d_v, queries y llaves de dimensión d_k, se realiza el producto punto de los queries con todas las llaves y luego eso se divide por la raíz cuadrada de d_k y se le aplica un softmax para obtener los respectivos pesos de los valores. Usualmente se aplica usando matrices Q, K, V para calcular la atención. Se puede ver con esta ecuación:

   $$Attention(Q,\ K,\ V)\ =\ softmax(\frac{QK^T}{\sqrt{d_k}})V$$

3. **¿Por qué se utiliza la atención de múltiples cabezales en Transformer?**

   La atención de múltiples cabezales resulta útil ya que en vez de tener solamente 1 función de atención , es mejor tener varias proyecciones lineares de d_k, d_k y d_v dimensiones para las llaves, valores y queries. Y en cada una de estas se realiza la función de atención paralelamente dando resultados d_v-dimesionales. Concatenándolos finalmente y proyectándolos obtenemos los valores finales.

4. **¿Cómo se incorporan los positional encodings en el modelo Transformer?**

   Debido a que no se emplea recurrencia ni convolución, se debe inyectar información respecto a la posición relativa o absoluta del token, y ahí es donde un positional encoding se suma a los embeddings al fondo de la pila de los encoder-decoder, y debido que tienen la misma dimensión, se suman.

5. **¿Cuáles son algunas aplicaciones de la arquitectura Transformer más allá de la Machine Translation?**

   Se evaluó si el transformador puede generalizarse a otras tareas con el English Constituency Parsing, el cual es bastante difícil o representa reto ya que puede presentar restricciones estructurales y que la salida sea mayor en tamaño que la entrada. Los transformadores según los resultados del paper, tuvieron buen desempeño en el ECP, mientras que los modelos RNN sequence-to-sequence quedaron deficientes en esta tarea.

Universidad del Valle de Guatemala
Dep. Ciencias de La Computación
Deep Learning
01/09/2024

Referencias:

- Vaswani, Shazeer, Parmar, Uszkoreit, Jones, N. Gomez, Kaiser, Polosukhin (2017) *Attention Is All You Need.* From https://arxiv.org/pdf/1706.03762