Eva Lucero Pérez Salcedo A01568830
7th Semester
Artificial Emotions:Going beyond Artificial Intelligence
Instituto Tecnologico de Monterrey Campus Guadalajara
Course provided by: Mahdi Zaarei and Alejandro de León Languré
11th October 2023

Importance of Attention Mechanisms in Sequence-to-Sequence Models for visual 3D models

Abstract

This paper has the purpose to focus on the attention mechanisms used in sequence-to-sequence models for processing and generating visual 3D models. Here we look for the difficulties it can come across as well as their drawbacks as being a complex dimensionality to analize at first hand by a computer. We go deep on how the existence of attention mechanisms can significantly improve to understand these visual models.
We go through literature review focused on attention mechanisms applied to 3D modeling, some databases like arXiv and papers to understand the key importance of attention mechanisms and their bases, as well as a deep dive into being used in natural language processing, fundamentally exploring their different types and how can they be computationally functional and their challenges while being implemented, emphasizing their importance in advancing seq2seq models for better handling of longer and more complex sequences.

INTRODUCTION

Sequence-to-sequence models commonly consist of a encoder-decoder method, where natural language processing takes part. The encoder works as a processing input for a decoder to generate an output sequence. Knowing this, the architecture models often used are RNNs, LSTMs or GRUs to handle the data for processing. Here we focus on the attention mechanisms that make traditional seq2seq models to allow a different approach and attention (as the name goes) to determine different parts of the input sequence to follow certain steps to attain the output generation, getting this model to have the ability to process longer and complex sequences that were limited by the traditional ones.

History of attention mechanisms

The attention mechanim was implemented to improve encoder-decoder models, it utilizes the most relevant input data in a sequence by weighting the most relevant vectors with higher values, by this we have a focused mechanism that can process high lenghted sequences into an output by the decoder.

> *“The attention mechanism was introduced by Bahdanau et al. (2014) to address the bottleneck problem that arises with the use of a fixed-length encoding vector, where the decoder would have limited access to the information provided by the input. This is thought to become especially problematic for long and/or complex sequences, where the dimensionality of their representation would be forced to be the same as for shorter or simpler sequences.” [8]*

As previous mentioned, Bahdanau et al.’s attention mechanism is divided into three computations to help a better understanding, this being alignment scores, weights and context vectors. A brief context about these computations are; In alignment scores, determines how much focus should be given to each part of the input while generating a specific and importance part of the output, they are calculated for each pair of input and output element to measure the compatibility between a particular time step and the output in their current state. Weights, as they come across attention weights, while the alignment scores are done the weights are normalized into a softmax function to make sure the weights from all the input steps sum up to 1, forming a probability distribution to indicate the importance of each input element in generating the next output, as the higher weights would have more influence. Context vectors, they are constructed for each output time step, this is calculated by the weighted sum of the input, usually in hidden stats. The context vector, which is a dynamic and adaptive representation of the input sequence, is then used along with the decoder's previous hidden state to generate the next output element to enable the model to focus on parts of the input, this way it can handle complex sequences more effective. These components working together make attention mechanisms to focus and adapt to the context of sequence-to-sequence models.

Fundamentals of Attention Mechanisms

The general attention mechanism uses three main components, queries (Q), keys (K) and values (V). Then it goes through this computational functions: [8]

Each query vector, $q = s_{t-1}$, is matched against a database of keys to compute a score value. This matching operation is computed as the dot product of the specific query under consideration with each key vector,k_i:

$$e_{q_2 k_i} = q * k_i$$

Then the scores are passed through a softmax operation to generate the weights:

$$\propto_{q_2 k_i} = softmax(e_{q_2 k_i})$$

The generalized attention is then computed by a weighted sum of the value vectors, V_{k_i}, where each value vector is paired with a corresponding key:

$$attention(q, K, V) = \Sigma_i \propto_{q_2 k_1} V_{k_i}$$

The generalized attention mechanism in neural networks works when it looks at a series of words, it picks a particular word and creates a special "query" related to that word. It then compares this query with "keys" that represent all the other words in the sequence. This comparison helps the mechanism understand how the chosen word is connected to the rest of the words. Next, it calculates the attention weights based on these comparisons. These weights help the mechanism decide which words are most important in relation to the chosen word. It then uses these weights to focus more on the important words and have less impact on the others. Finally, it creates an attention output for the chosen word. This output is a new representation of the word that reflects its relationship and importance to the other words in the sequence.

Types of Attention Mechanisms

Attention mechanisms in computer vision are categorized based on their data domain into six main types, these being:

Channel Attention: Focuses on what to pay attention to within the channel domain.

Spatial Attention: Concentrates on where to pay attention, dealing with spatial regions of the input.

Temporal Attention: Addresses when important data occurs, particularly relevant in sequences or time-dependent data.

Branch Attention: Determines which branch or path to focus on in a network.

Channel & Spatial Attention: A hybrid category combining aspects of both channel and spatial attention.

Spatial & Temporal Attention: Another hybrid category that combines the spatial and temporal attention. [6]

Case Studies and Applications

Attention Mechanisms in 3D Point Cloud Object Detection

The study examines the effectiveness of various attention mechanisms (both 2D and 3D) in the context of 3D point cloud object detection. The primary focus is on understanding which types of attention mechanisms are most suitable for this task. Experiments are conducted using the SUN RGB-D and ScanNetV2 datasets. Various attention modules are tested, including classical 2D attentions (e.g., Non-local, SE, CBAM) and novel 3D attentions (e.g., Point-Attention, Point Transformer).

It is noted that compact attention structures like SE (Squeeze-and-Excitation) and CBAM (Convolutional Block Attention Module) are effective and efficient for 3D point cloud feature refinement. The paper discusses the complexity of different attention modules when integrated into the VoteNet backbone, considering factors like model size, training time, and inference time. [7]

The study also finds that channel-related information is more crucial than spatial information in attention modules for point cloud feature representations.
The Point Transformer shows significant effectiveness, particularly in handling complex point cloud scenes.

CONCLUSION

The exploration of attention mechanisms in sequence-to-sequence models, especially in the realm of visual 3D modeling, has revealed their indispensable role in modern computational vision and natural language processing. Attention mechanisms, by enabling models to focus selectively on parts of a sequence, overcome limitations in earlier seq2seq models, particularly in handling lengthy and complex input sequences. This selective focus translates into improved model performance, enhanced understanding of context, and more accurate predictions, especially in high-dimensional data environments like 3D modeling.

In the specific context of 3D point cloud object detection, attention mechanisms have demonstrated their capability to refine features more effectively, with particular emphasis on channel-related information over spatial details. Compact structures like SE and CBAM, alongside novel implementations like the Point Transformer, have shown promising results in managing the intricacies of 3D point cloud scenes.

In conclusion, attention mechanisms represent a significant advancement in the field of machine learning, offering versatile and powerful tools for enhancing the capabilities of seq2seq models. Their ability to adapt to complex sequences and provide context-aware processing makes them invaluable for a wide range of applications, from natural language processing to advanced computer vision tasks involving 3D modeling. As research in this area continues to evolve, we can anticipate even more sophisticated and efficient models capable of overcoming more limitations in analyzing and interpreting complex data structures.

In summary, the attention mechanism in seq2seq models marks a crucial evolution, enabling more complex and context-aware processing in various NLP tasks. It has significantly improved the performance of seq2seq models, especially in applications requiring an understanding of long or complex input sequences.

This paper provides a comprehensive analysis of how different attention mechanisms influence 3D point cloud object detection, offering valuable insights and standards for future research in this area.

References

[1]I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 3104–3112.

[2]D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473, 2014.

[3]A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017, pp. 5998–6008.

[4]I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[5]D. Jurafsky and J. H. Martin, Speech and Language Processing. 3rd ed., Draft, 2020.

[6]M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu, "Attention mechanisms in computer vision: A survey," Comput. Visual Media, vol. 8, no. 3, pp. 331–368, Sep. 2022.

[7]S. Qiu, Y. Wu, S. Anwar, and C. Li, "Investigating Attention Mechanism in 3D Point Cloud Object Detection," arXiv preprint arXiv:2108.00620v2, Oct. 14, 2021.

[8]J. Brownlee, "The Attention Mechanism From Scratch," Machine Learning Mastery. [Online]. Available: https://machinelearningmastery.com/the-attention-mechanism-from-scratch/. [Accessed: Oct. 11, 2023].