JASTON STEPHEN NGADATA

RU/BSCSE/2023/041

ASSIGNMENT ONE

1. SUMMARY OF THE INTRODUCTION

The paper introduces a new neural network architecture called the Transformer, designed for sequence-to-sequence tasks such as machine translation. Traditional models like RNNs and LSTMs process words sequentially, which limits speed and struggles with long-term dependencies. Earlier improvements like attention mechanisms were added to such models but still depended on recurrence. The authors argue that recurrence is not necessary, and instead propose a model based solely on attention mechanisms, allowing for parallel processing of entire sequences. This results in significantly faster training while improving accuracy on translation tasks.

The introduction highlights that natural language processing tasks require capturing relationships between words regardless of their distance. The Transformer accomplishes this using self-attention, enabling the model to understand global dependencies without sequential computation.

2. Name of the Paper and Author(s)

Paper Title: “Attention Is All You Need”

Authors: Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones,

Aidan N. Gomez, Ł ukasz Kaiser, and Illia Polosukhin.

Published: 2017 (NeurIPS Conference)

3. SUMMARY OF THE PROBLEM STATEMENT

the authors identify the following main problems with previous sequence models (rnns, grus, lstms):

i. Sequential processing slows training.

RNN-based architectures read tokens one at a time, preventing the use of parallel computation and making training slow and expensive.

ii. Difficulty learning long-range dependencies.

Even advanced models struggle when important words are far apart in a sentence.

iii. High computational cost for long sequences.

Attention layers in RNNs require computation proportional to sequence length.

iv. Translation quality is limited.

Existing architectures show performance bottlenecks, especially for long or complex sentences.

In short, the problem is:

Existing sequence-to-sequence models are slow, computationally expensive, and limited in understanding long-distance relationships in text.

4. PROVIDE THE OBJECTIVES

the main objectives of the paper were:

i. To replace recurrent neural networks with a fully attention-based architecture.

The authors aim to show that recurrence is not necessary for sequence modeling.

ii. To improve computational efficiency.

The goal is to design a model that trains faster by leveraging parallelization.

iv. To improve translation accuracy.

The Transformer is intended to outperform RNN/LSTM models on machine translation benchmarks.

v. To model long-range dependencies better.

The self-attention mechanism is designed to capture global relationships between all words in a sequence.

vi. To simplify model architecture.

By removing recurrence, the authors aim for a simpler and more scalable model.

5. SUMMARY OF THE PROBLEM STATEMENT

The core problem addressed is that previous neural sequence models depend on recurrence, which limits their ability to process input efficiently and capture distant dependencies. These limitations slow down training and reduce translation performance. The authors propose a new architecture addressing these issues using only attention mechanisms.

6. PROVIDE THE GAP OF THE PAPER

the research gap identified in the paper is:

i. Lack of models using attention without recurrence.

Before this work, attention was only used as an additional component on top of RNNs or CNNs, not as the primary mechanism.

ii. Inefficiency in handling long sequences.

No previous architecture provided a fast, parallel way to capture long-range dependencies.

iii. Limited exploration of self-attention as a replacement for RNNs.

Self-attention had not been fully studied as a standalone architecture for sequence-to-sequence tasks.

iv. Scalability issues in previous models.

Traditional models struggled with large datasets and long sentences due to computational constraints.