Name:- P.sumanth
Reg.No:-9922005061

:-ASSIGNMENT-2:-

summary of "Attention IS ALL YOU NEED"(NIPS 2017)

Authors:-

Ashish vaswani, Noam shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N.Gomez, Lukasz kaiser and illia Polosukhin.

Introduction:-

* The Paper introduces "the transformer". a novel neural network architecture for sequence transduction (e.g; Machine translation).
* Unlike previous models that rely on "Recurrent Neural Networks(RNNs)" or "Convolutional Neural Networks (CNNs)" the transformer is based entirely on "self-attention mechanisms".
* This removes the need for recurrence, enabling better parallelization and faster training while achieving superior performance.
* The transformer achieves "state-of-the-art results" in machine translation tasks with significantly lower computational costs.

Key contributions:-

1) Self-Attention Mechanism:- It replaces recurrence, allowing models to capture long range dependencies more effectively.
2) Multi-Head Attention: enables the model to focus on different Parts of the input simultaneously.

3) Positional Encoding compensates for the lack of sequential structure in self-attention.

4) Layer Normalization & Residual connections improve training stability.

5) Parallelized computation significantly reduces training time compared to RNN-based models.

Model Architecture:- The Transformer:-

The transformer follows the encoder-decoder architecture:-

* Encoder:- Maps input sequences to a continuous representation.

* Decoder:- Generates the output sequence step by step.

Each encoder and decoder block consists of:-

1) Multi-Head self-Attention:- captures dependencies between woords.

2) feed-forward Network:- A Position-wise fully connected netwoork.

3) Layer Noormalization & Residual connections:- improve gradient flow and stability.

Key Innovations:-

* Scaled Dot-Product Attention:- computes attention scoores efficiently.

* Multi-Head Attention:- uses multiple attention heads to capture diverse feautures.

Positional Encoding:- Injects order information into the model since self-attention lacks sequential dependencies

Advantages of the Transformer:-

* Higher efficiency:- fully parallelized training compared to sequential RNNs.

* Better long-range dependency modeling:- Self-attention allows direct connections b/w distant words.

* Reduced training cost:- Requires significantly fewer resources compared to RNNs and CNNs.

* State-of-the-art performance:- out performs previous models in machine translation tasks.

Results & Performance:-

* Achieved 28.4 BLEU on WMT 2014 English-to-German translation, surpassing previous state-of-the-art models.

* Achieved 41.8 BLEU on WMT 2014 English-to-french translation with just 3.5 days of training on 8 GPUs.

* out performed RNN-based models in English Constituency Parsing

Conclusion & future work:-

* The transformer eliminates the need for recurrence, revolutionizing sequence modeling.
* future research directions include applying self-attention to other domains like images, audio and video.
* The Transformer has since influenced advancements like BERT, GPT and T5, shaping the future of NLP.

This paper laid the foundation for modern NLP models and self-attention-based architectures.