Content Understood

* In this article, the Transformer architecture revolutionizes sequence-to-sequence translations in natural language processing.
* One can do away with recurrent or convolutional layers when using a transformer as they rely solely on self-attention strategies.
* This structure enables processing an entire sentence as a whole, thereby capturing actual contexts and long distance dependencies.
* For example, in the field of data processing, it is often the case that while we are using outdated models for our recurrent neural networks like RNNs; as a result, we lack both efficiency and parallelism.
* The authors reached the highest BLEU scores in WMT 2014 English-German, English-French by experimenting with data extensively.
* The Transformer requires less learning time than the previous models because of its parallel operation which has helped it achieve a higher accuracy.
* The clarity of the model is enhanced as attention heads within the model focus on various aspects of the input data.
* Altogether with Transformer, natural language processing has moved forward in great measure bringing about numerous possibilities for us both in terms of research and real-world usage.

Content Not Understood

* The self-attention model has several other refined aspects including the way in which attention scores are computed and influence word embeddings.
* It is not simple precisely what is involved. Especially difficult is the problem of deciding which features the model needs to pay attention to and how this affects the overall performance.
* It is established that the Transformer has to be changed in such a way so it could be used in various domains, apart from the original ones such as pictures, audio and video.
* I desire it so much that deeper insights are achieved regarding what drives an all important aspect behind recognition models.
* Key issues in the field continue to be lexical ambiguity and machine translation (MT).