The core idea behind the self-attention mechanism, as proposed in the paper "Attention Is All You Need," is to compute a representation of a sequence by relating different positions within that sequence. Self-attention, also known as intra-attention, allows the model to weigh the importance of each position in the sequence relative to others, enabling it to capture dependencies between distant positions efficiently.  
  
In traditional sequence models, capturing long-range dependencies can be computationally expensive and challenging. However, self-attention reduces this complexity to a constant number of operations, allowing the model to focus on relevant parts of the sequence regardless of their distance from each other. This is achieved by calculating attention scores for each position, which are then used to create a weighted sum of the input features, effectively highlighting important parts of the sequence.  
  
The paper addresses a potential downside of self-attention, which is the reduced effective resolution due to averaging attention-weighted positions. This issue is mitigated through the use of Multi-Head Attention, which allows the model to attend to information from different representation subspaces at different positions, thereby enhancing its ability to capture complex patterns and relationships within the data.  
  
Self-attention has been successfully applied to various tasks, including reading comprehension, abstractive summarization, textual entailment, and learning task-independent sentence representations, demonstrating its versatility and effectiveness in processing sequential data.