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. This mechanism, also known as intra-attention, allows the model to focus on different parts of the input sequence when constructing its output representation.  
  
Self-attention operates by assigning attention scores to each position in the sequence relative to every other position. These scores determine how much influence each position should have in the final representation of a particular position. The attention scores are typically computed using a compatibility function, such as the dot product, followed by a softmax operation to ensure they sum to one.  
  
One of the key advantages of self-attention is its ability to capture dependencies between distant positions in the sequence efficiently. In traditional sequence models, such as recurrent neural networks (RNNs), capturing long-range dependencies can be computationally expensive and challenging. In contrast, self-attention reduces this complexity to a constant number of operations, regardless of the distance between positions.  
  
However, this efficiency comes at the cost of reduced effective resolution due to the averaging of attention-weighted positions. To address this, the paper introduces Multi-Head Attention, which allows the model to attend to information from multiple representation subspaces simultaneously, thereby enhancing the model's ability to capture complex dependencies.  
  
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 natural language processing.