This document explains the attention mechanism used in modern deep learning models. Attention allows neural networks to dynamically focus on the most relevant parts of the input data. It significantly improves performance in tasks such as machine translation, text summarization, and question answering. Self-attention is the core idea behind Transformer models like BERT and GPT, enabling parallel processing and better context understanding.