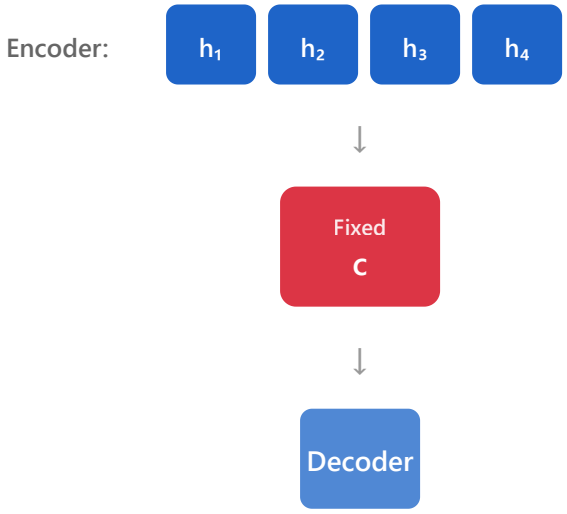
Core Ideas of Attention Mechanism

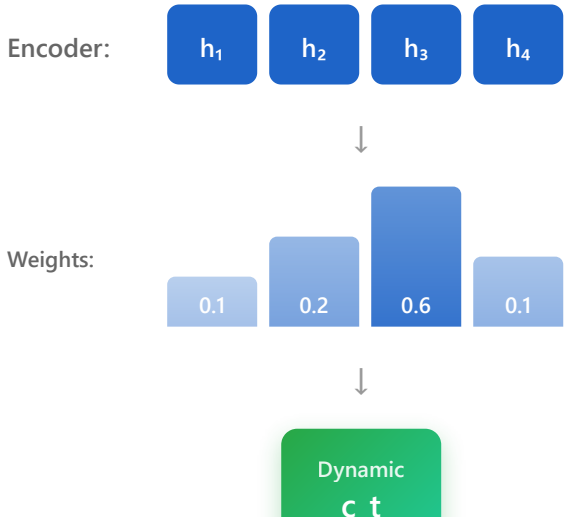
💡 Central Concept

Instead of compressing everything into a single context vector, **let the decoder selectively focus on different parts of the input** at each time step based on what's most relevant.

Without Attention



With Attention



Same static context for all decoder steps. Information bottleneck. Long sequences lose details.

Dynamic context computed at each step. Weighted focus on relevant parts. Full input information accessible.

🎯 Key Benefits of Attention

- ✓ No information bottleneck
- ✓ Handles long sequences
- ✓ Selective focus
- ✓ Better alignment
- ✓ Interpretable weights
- ✓ Improved performance