Types of Unconditioned Prediction

Exploring different approaches to sequence prediction in machine learning, each offering unique trade-offs between context understanding, computational efficiency, and prediction capabilities.

Understanding the Visualizations

- Available Context: Information the model can access
- Current Position: Token being predicted
- Unavailable Context: Information not used

These visualizations show how different prediction models access and use context. Blue squares represent available information, red squares show the current prediction position, and gray squares indicate information the model cannot access.

Left-to-right Autoregressive Prediction

$$P(X) = \prod_{i=1}^{|X|} P(x_i | x_1, ..., x_{i-1})$$

Models text understanding similarly to human reading, progressively building context from left to right. Each prediction incorporates all previous information, creating a rich but sequential understanding process.

Examples

RNN, Transformer LM

Complexity

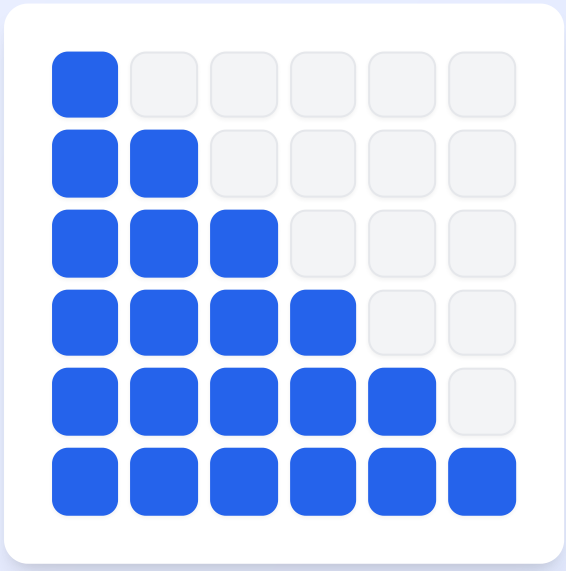
O(n) for generation, where n is sequence length

Technical Details

The model processes input sequentially, allowing it to maintain a thorough understanding of context. This approach mirrors how humans read and understand text, making it particularly effective for tasks requiring deep contextual awareness.

Key Limitations

Cannot utilize future context, sequential generation can be slow



The blue squares show how context accumulates from left to right, like reading a book word by word

Left-to-right Markov Chain (order n-1)

$$P(X) = \prod_{i=1}^{|X|} P(x_i | x_{i-n+1}, \ldots, x_{i-1})$$

Focuses on recent context using a sliding window approach. Like having a limited short-term memory, it trades comprehensive understanding for computational efficiency.

Examples

n-gram LM, feed-forward LM

Complexity

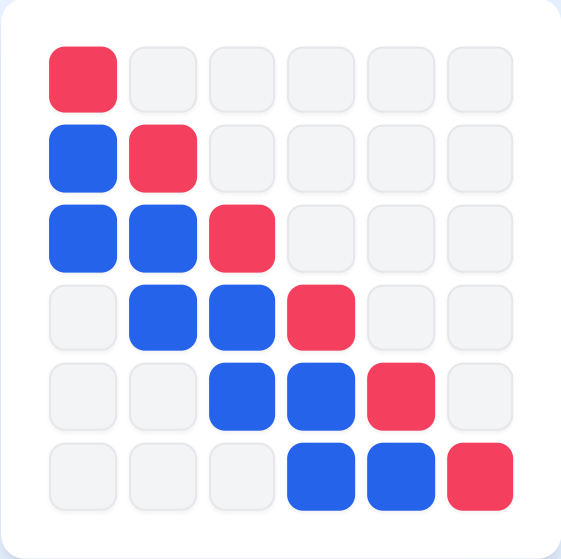
O(1) per token, fixed context window

Technical Details

By considering only a fixed window of previous tokens, this approach balances context awareness with processing speed. While it may miss long-range patterns, it excels at capturing local dependencies.

Key Limitations

Fixed context window may miss important long-range patterns



The red square shows the current prediction, using only the recent blue context squares

Independent Prediction

$$P(X) = \prod_{i=1}^{|X|} P(x_i)$$

Makes predictions for each position independently, like trying to guess words without reading the rest of the sentence. Fast but limited in understanding context.

Examples

Unigram model

Complexity

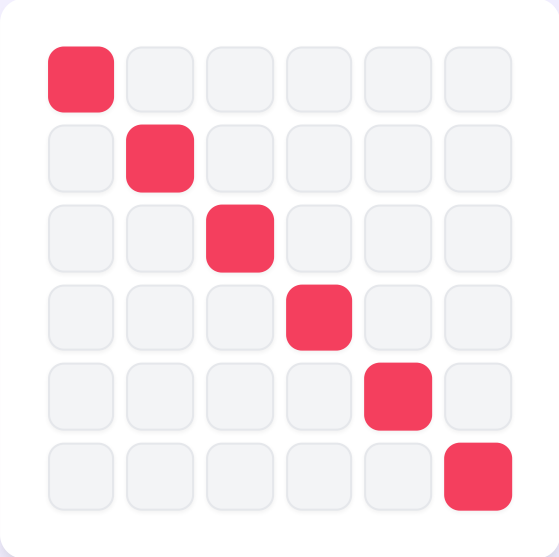
O(1) per token, fastest prediction

Technical Details

The simplest approach, treating each token as independent. While this limits understanding of relationships between tokens, it offers maximum parallelization and computational efficiency.

Key Limitations

Ignores all contextual relationships between tokens



Each red square makes predictions independently, without using information from other positions

Bidirectional Prediction

$$P(X) \neq \prod_{i=1}^{|X|} P(x_i | x_{\neq i})$$

Utilizes both past and future context, like having the full picture before making any predictions. This comprehensive approach enables rich understanding but requires special training methods.

Examples

Masked language model

Complexity

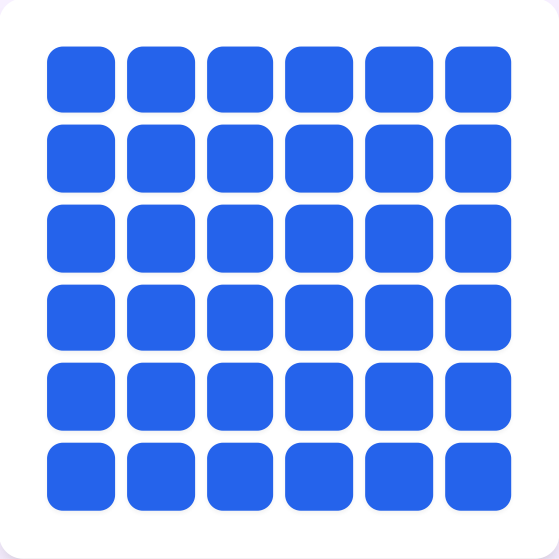
O(n) for attention-based models

Technical Details

By considering both past and future context, this approach achieves the most complete understanding. However, it requires specialized training techniques and cannot be used for standard sequence generation.

Key Limitations

Cannot be used for direct sequence generation



All squares are connected, showing how the model can use both past and future context