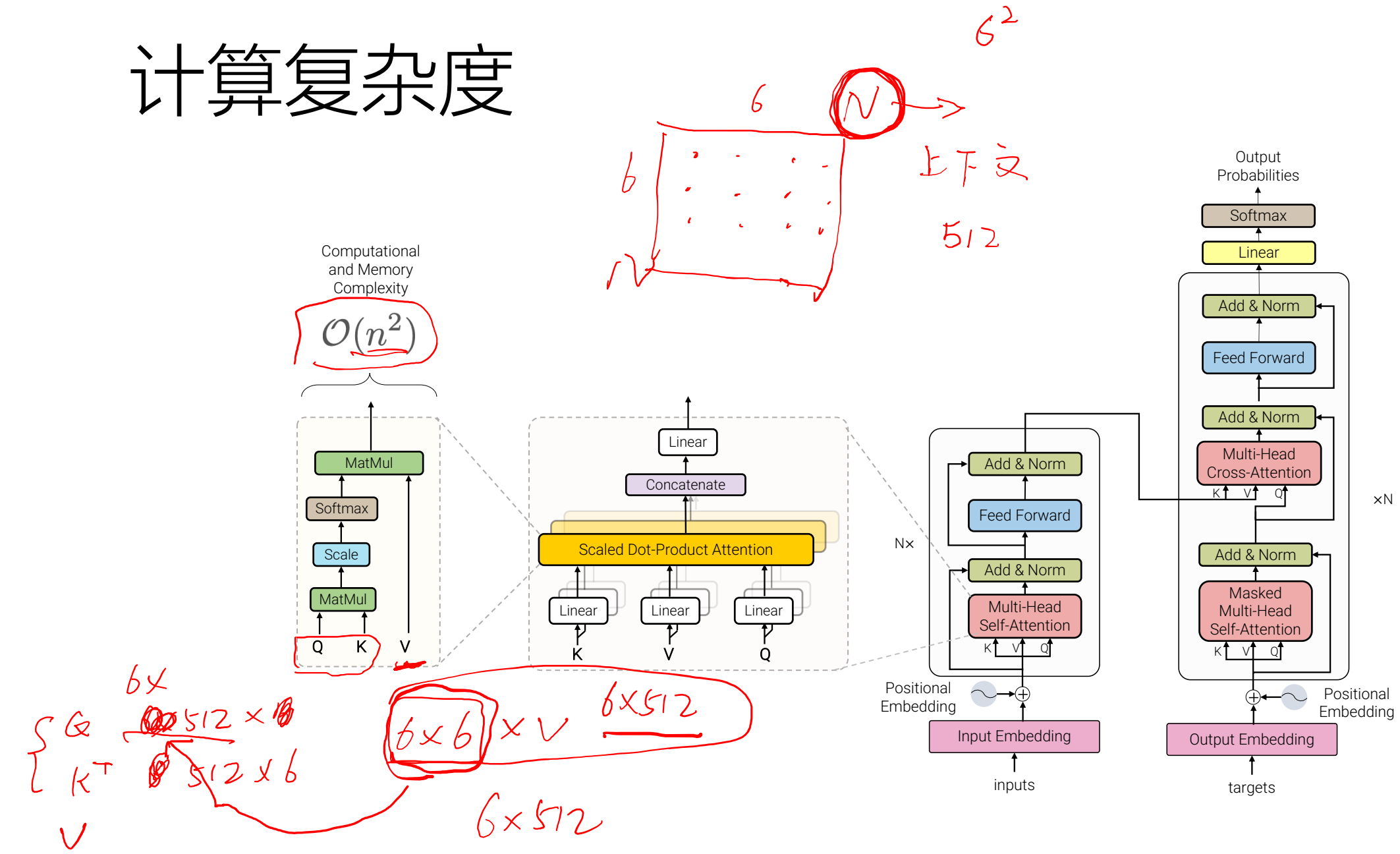
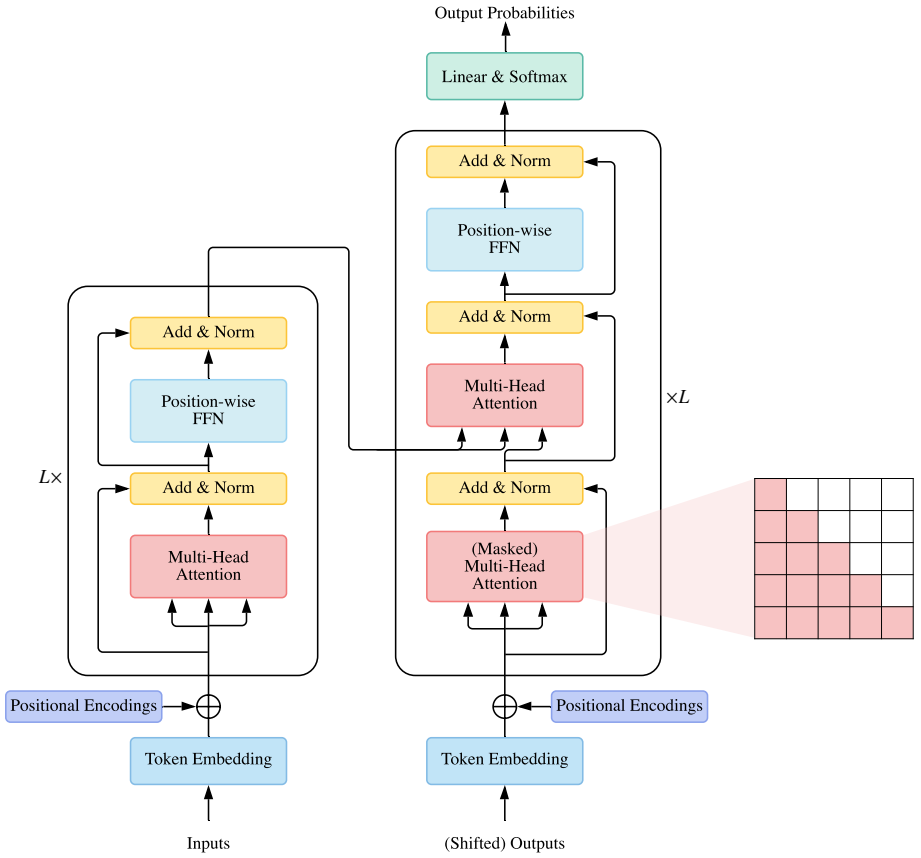
Transformer架构优化

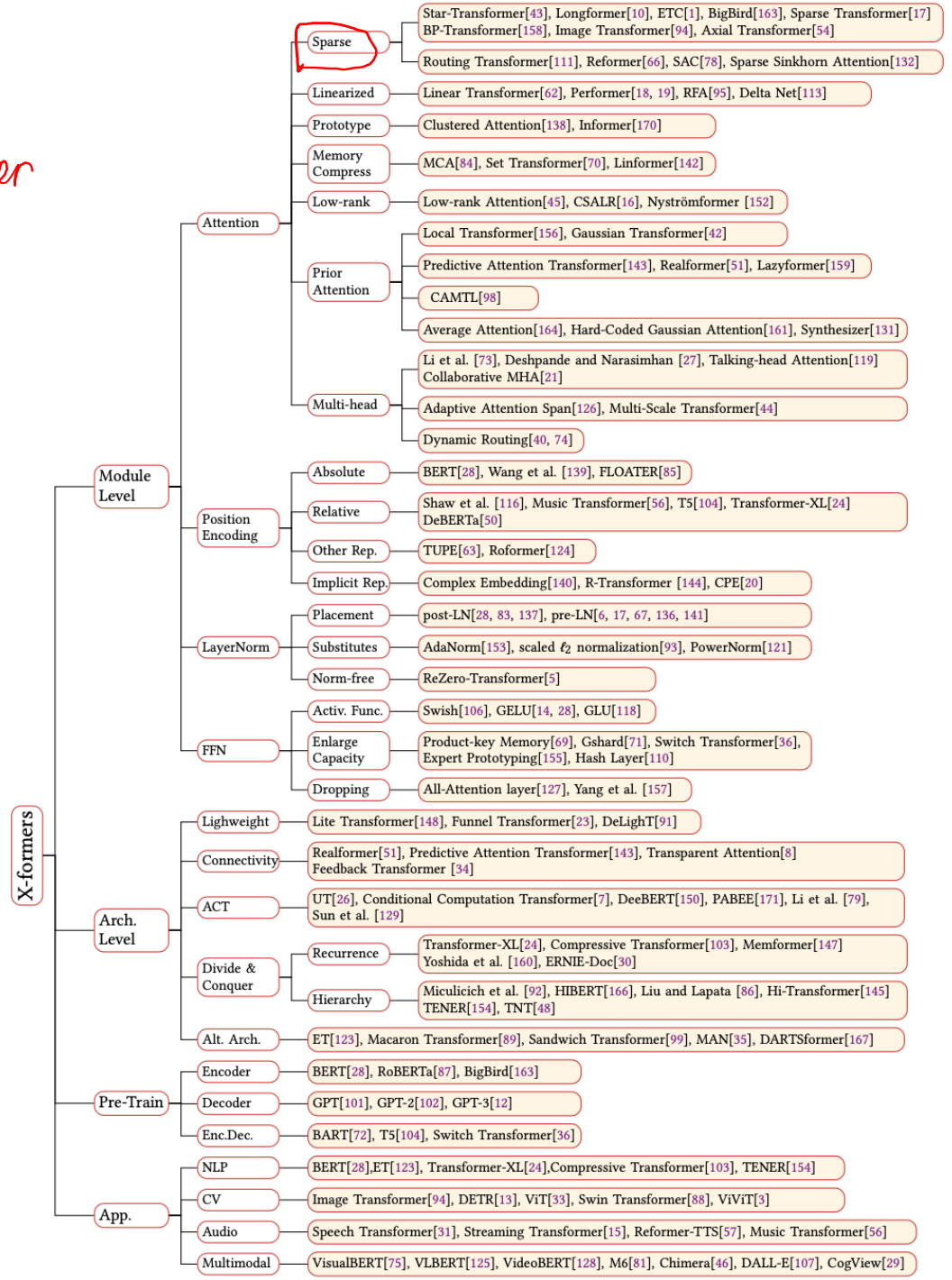
Scott

计算复杂度


Computational and Memory Complexity
$\mathcal{O}(n^2)$
MatMul
Softmax
Scale
MatMul
Q
K
V
Linear
Concatenate
Scaled Dot-Product Attention
Linear
Linear
Linear
K
V
Q
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Self-Attention
K
V
Q
Positional Embedding
Input Embedding
inputs
Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Cross-Attention
K
V
Q
Add & Norm
Masked Multi-Head Self-Attention
K
V
Q
×N
Positional Embedding
Output Embedding
targets

Transformer家族 → x-former

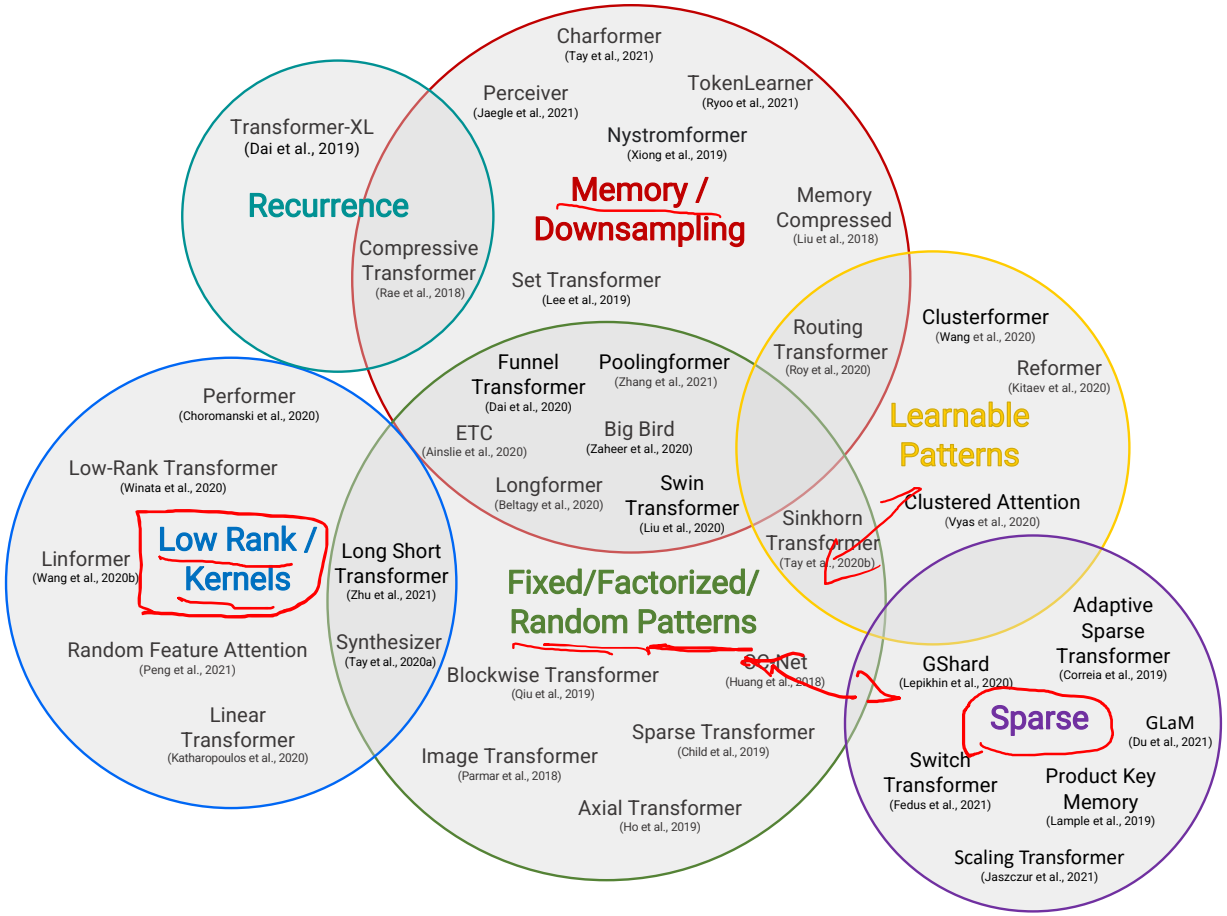




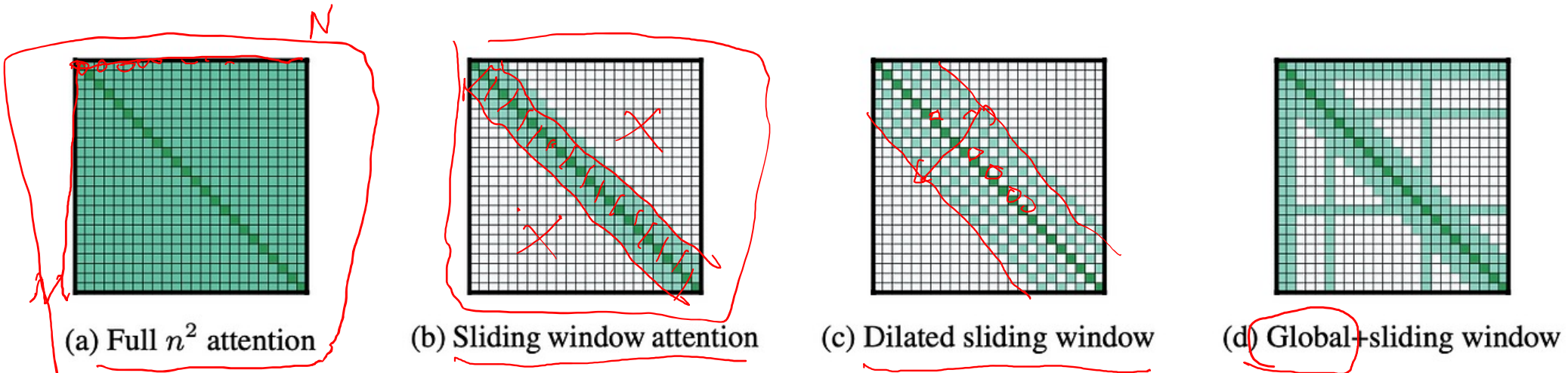
高效Transformer架构设计

Efficient Transformer

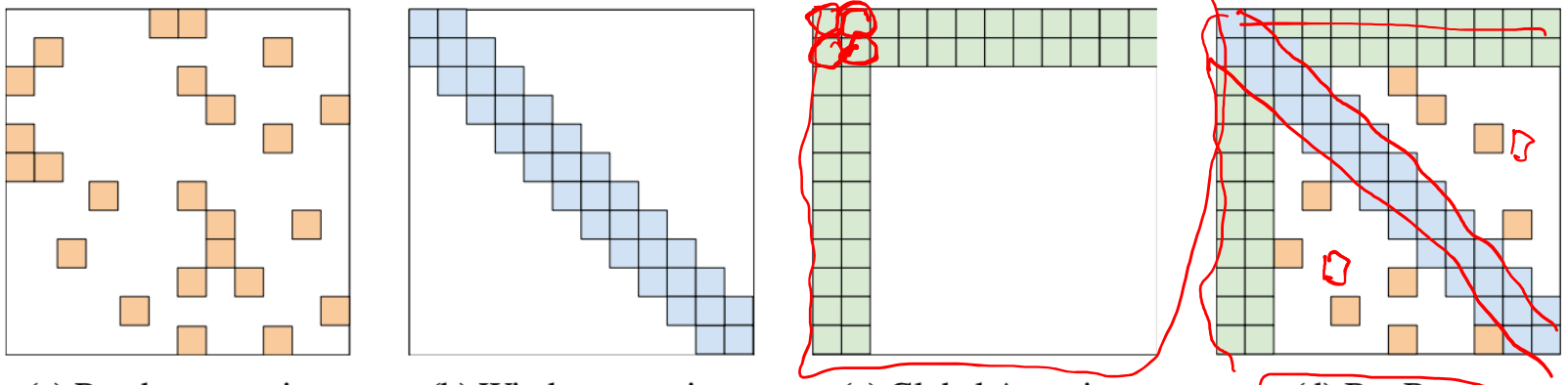
- 增加上下文长度
- 提高速度
- 减少内存使用



稀疏方法

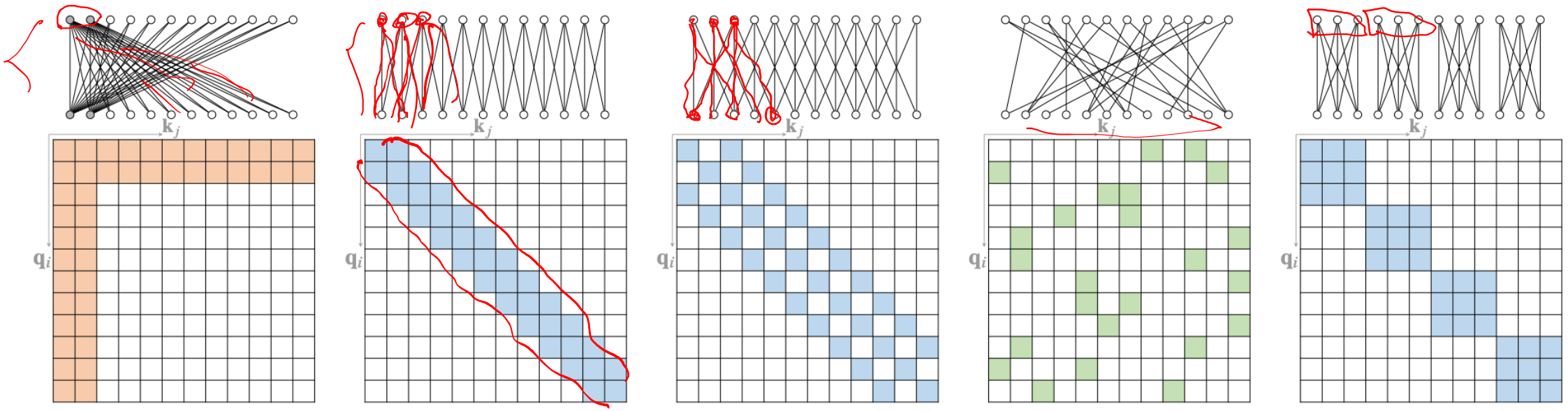


(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window



(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

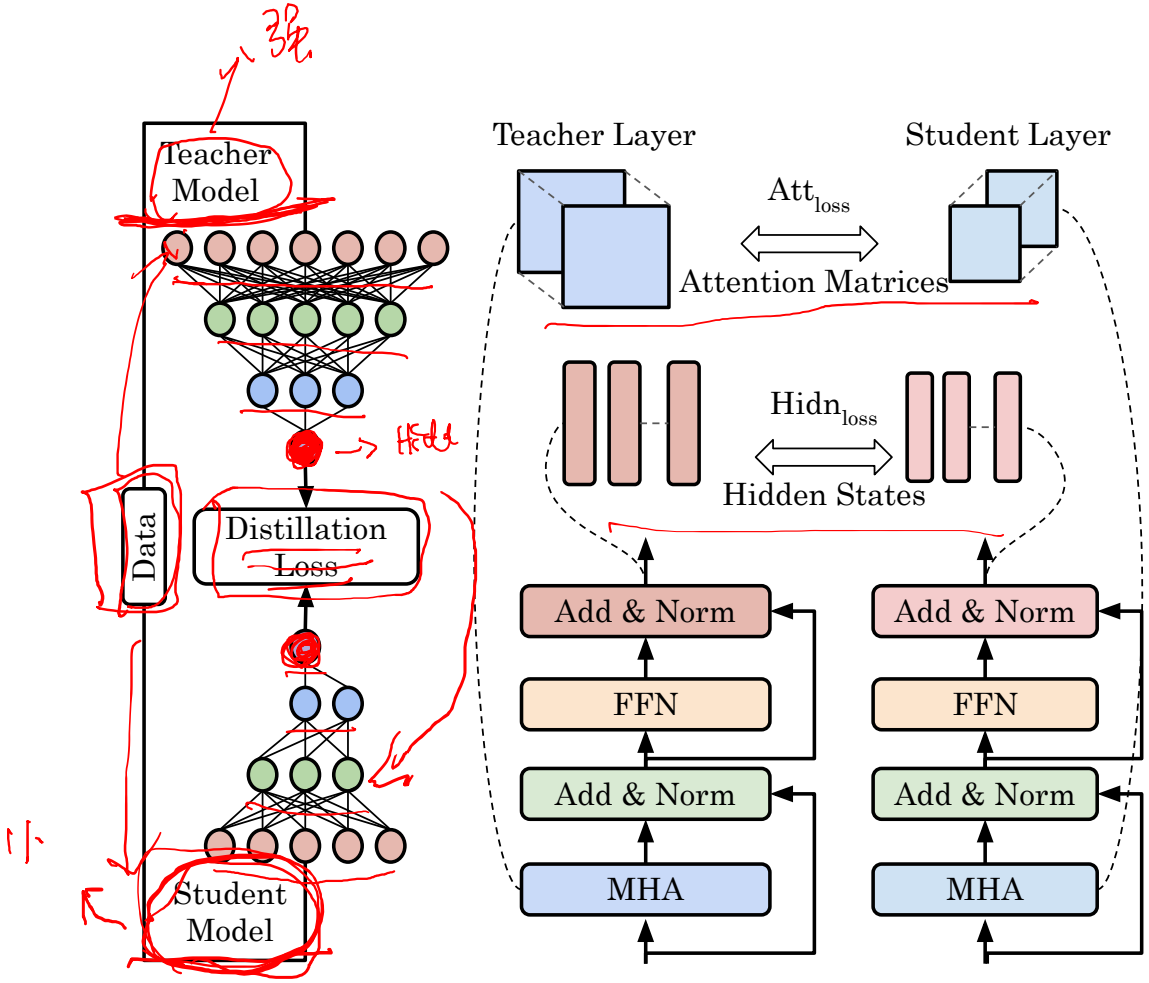
稀疏方法



(a) global (b) band (c) dilated (d) random (e) block local

推理与压缩

- 剪枝
- 蒸馏
- 量化



(a) KD Overview (b) Transformer Distillation

Socratic Method

- https://princeton-nlp.github.io/SocraticAI/

- CRITIC: LARGE LANGUAGE MODELS CAN SELFCORRECT WITH TOOL-INTERACTIVE CRITIQUING
https://arxiv.org/pdf/2305.11738.pdf
- INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback
https://arxiv.org/abs/2305.14282
- Pinpoint, Not Criticize: Refining Large Language Models via Fine-Grained Actionable Feedback
- Prompting Large Language Models With the Socratic Method
https://arxiv.org/pdf/2303.08769.pdf
- SOCREVAL: Large Language Models with the Socratic Method for Reference-Free Reasoning Evaluation
https://arxiv.org/pdf/2310.00074.pdf
- SocraSynth: Socratic Synthesis for Reasoning and Decision Making
https://www.researchgate.net/publication/373753725_SocraSynth_Socratic_Synthesis_for_Reasoning_and_Decision_Making