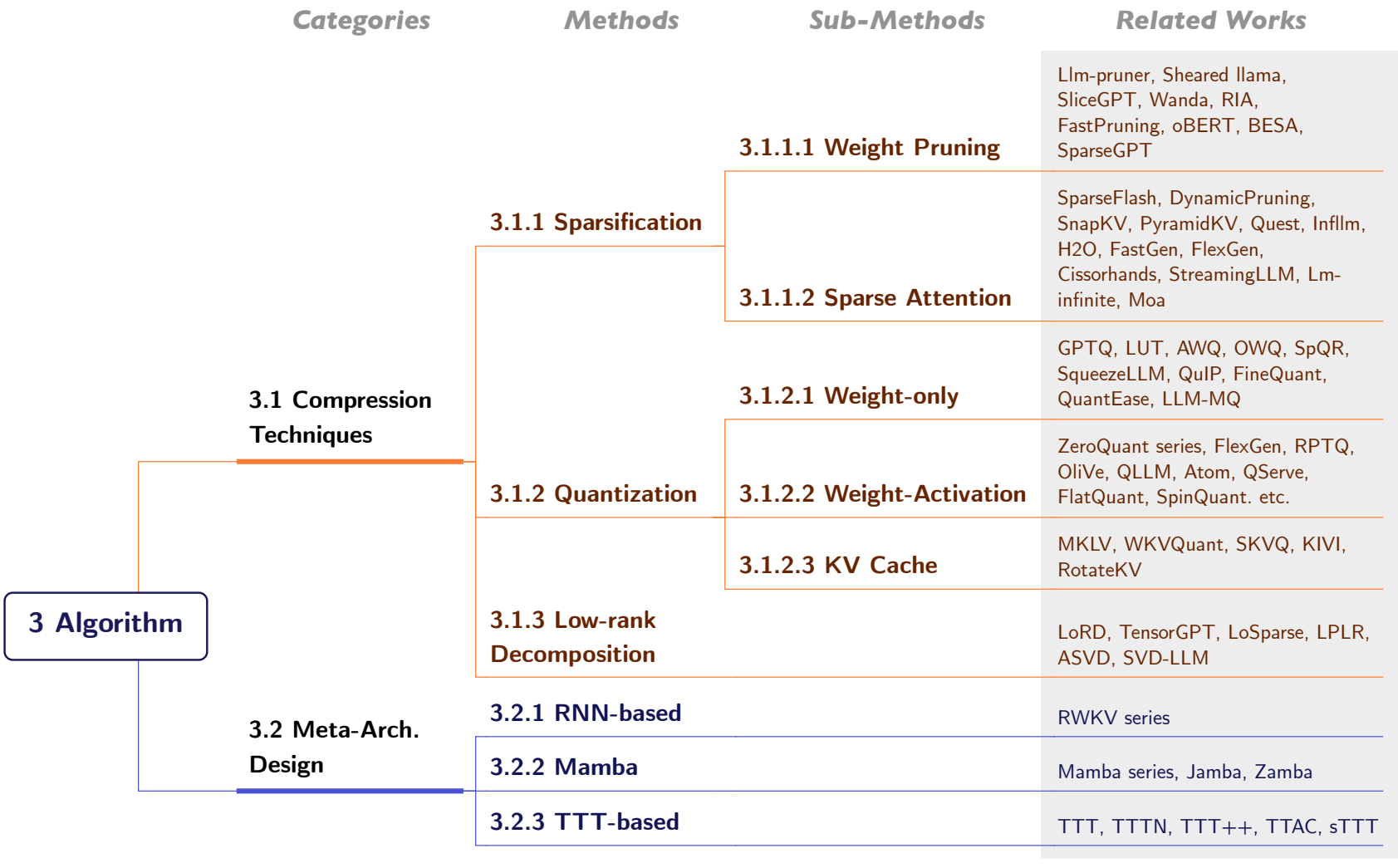
Categories
Methods
Sub-Methods
Related Works
3 Algorithm
3.1 Compression Techniques
3.2 Meta-Arch. Design
3.1.1 Sparsification
3.1.2 Quantization
3.1.3 Low-rank Decomposition
3.2.1 RNN-based
3.2.2 Mamba
3.2.3 TTT-based
3.1.1.1 Weight Pruning
3.1.1.2 Sparse Attention
3.1.2.1 Weight-only
3.1.2.2 Weight-Activation
3.1.2.3 KV Cache
Llm-pruner, Sheared llama, SliceGPT, Wanda, RIA, FastPruning, oBERT, BESA, SparseGPT
SparseFlash, DynamicPruning, SnapKV, PyramidKV, Quest, Infllm, H2O, FastGen, FlexGen, Cissorhands, StreamingLLM, Lm-infinite, Moa
GPTQ, LUT, AWQ, OWQ, SpQR, SqueezeLLM, QuIP, FineQuant, QuantEase, LLM-MQ
ZeroQuant series, FlexGen, RPTQ, OliVe, QLLM, Atom, QServe, FlatQuant, SpinQuant. etc.
MKLV, WKVQuant, SKVQ, KIVI, RotateKV
LoRD, TensorGPT, LoSparse, LPLR, ASVD, SVD-LLM
RWKV series
Mamba series, Jamba, Zamba
TTT, TTTN, TTT++, TTAC, sTTT