Fairness principle: The performance advantage of a pruning paper should originate from the pruning method stage, not pretraining or finetuning. In terms of strictness, S1 < S2 < S3 < S4. S2 (the most popular one) cannot ensure fairness. Comparison setup **≥S3** is suggested to ensure fairness.

S1: Given the same

Dataset
Network structure
Total sparsity/speedup

S2: S1 + Keep the same base model (M1 arises at this setup)

S1-S4: Four pruning comparison setups
M1-M2: Two mysteries in filter pruning

Sparsifying action

S3: S2 + Keep the same finetuning

🔍 Unveil M1: Network trainability is damaged by the sparsifying action, slowing down the finetuning optimization; larger LR makes the model converge faster, thus better performance is observed earlier. The performance is not really "boosted".

| Random model (big dense) | Pretraining (big dense) | Base model (big dense) | Pruning method | Pruned model (small dense) | Finetuning (small dense) | Final model (small dense) |

Pruning pipeline

M2: Training the pruned model from scratch with adjusted (typically prolonged) epochs can produce the final model with similar performance.

M2 holds under S3, not hold under S2. To unveil M2, we need to unveil M1 first.

M1: A larger finetuning LR can significantly "boost" the final performance.

Scratch training pipeline

| Pruned model (small dense) | Training with adjusted epochs (small dense) | Final model (small dense) |

S4: S3 + Keep the same pruning cost