**FOLDING LEARNED NETWORKS INTO EXPLICITLY RECURRENT FORMS**

**Stage 1 LSM**
- Wide Residual Networks (WRNs) [7]
  - Conv $\otimes W$ $\rightarrow$ 1 layers
  - ResNet, ResNetV2 [9], DenseNet [11] all serve as additional comparison points

**Stage 2 LSM**
- Shared Wide Residual Networks (SWRNs)
  - $\otimes\otimes W$ $\rightarrow$ 4 parameter templates per group of layers
  - Groups consist of layers with the same parameters via template tensor
  - Setting $k = 1 \rightarrow 4$ yields one template per layer; denote such model as SWR-1

**Stage 3 LSM**
- Performance boost at one template per layer (i.e., without any parameter reduction)
- 2 templates per 6-layers (SWRN-XX-2): even better on CIFAR-10, competitive on CIFAR-100

**COMPARISON TO NEURAL ARCHITECTURE SEARCH**
- Faster training than recent gradient-based NAS, but...
  - less parameter efficient
- Non-NAS techniques (e.g., cutout, auxiliary towers, hyper-level optimisation) contribute to gains reported in many NAS results
- Cutout regularization boosts baseline WRN model... and on SWRN models (see Table to the right)

**EXTRAPOLATION BIAS**
- Challenge: Learn Shortest-Paths Algorithm:
  - Faster adaptation to harder examples
  - Better overall accuracy

**EVALUATION OF SHARING PATTERNS**
- Structure emerges quickly in the first epochs of CIFAR-10 training
- Folding possible after only 22% of total training iterations
- Faster evolution & stronger patterns via reparameterization

**IMPROVED PARAMETER EFFICIENCY**

**Baseline Models:**
- WRN-L $\rightarrow$ 2 layers
  - $\otimes\otimes W$ $\rightarrow$ 4 parameter templates per group of layers
  - Groups consist of layers with the same parameters via template tensor
  - Setting $k = 1 \rightarrow 4$ yields one template per layer; denote such model as SWR-1

**Our Models:**
- SWRN-L $\rightarrow 2$ layers
  - 4 parameter templates per group of layers
  - Groups consist of layers with the same parameters via template tensor
  - Setting $k = 1 \rightarrow 4$ yields one template per layer; denote such model as SWR-L

**Results:**
- Performance boost at one template per layer (i.e., without any parameter reduction)
- 2 templates per 6-layers (SWRN-XX-2): even better on CIFAR-10, competitive on CIFAR-100

**Experimental Results:**
- Classification: better accuracy and smaller CNNs
- Learning may discover explicitly recurrent architectures (new form of architecture search)
- Synthetic task: improved extrapolation ability

**Soft Parameter Sharing:**
- Trained networks present excessive redundancy
- Possible cause: different layers cannot re-use parameters
- Proposed method: layers learn to share parameters, alternatively seen as re-parametrization

**Approach:**
- Share parameters across layers in a neural network
- Soft sharing: relaxation of discrete optimization problem of selecting layers from a pool

**Consequences:**
- Decomposes parameter count from network depth
- Imposes inductive bias similar to RNNs
- Benefits optimization: accelerates learning of layers with similar functionality (redundant parameters $\rightarrow$ shared parameters)

**Parameters (M)**
- SWRN 28-2-1
- SWRN 28-4-2
- SWRN 28-10-2
- SWRN 28-16-2
- SWRN 28-18-2
- WRN 28-10
- WRN 40-4
- RNX 8x64
- RNX 16x64
- DN 40-12
- DN 100-24

**Test Error (%):**
- SWRN 28-2-1
- SWRN 28-4-2
- SWRN 28-10-2
- SWRN 28-16-2
- SWRN 28-18-2
- WRN 28-10
- WRN 40-4
- RNX 8x64
- RNX 16x64
- DN 40-12
- DN 100-24

**Soft Parameter Sharing**
- Trained networks present excessive redundancy
- Possible cause: different layers cannot re-use parameters
- Proposed method: layers learn to share parameters, alternatively seen as re-parametrization

**Approach:**
- Share parameters across layers in a neural network
- Soft sharing: relaxation of discrete optimization problem of selecting layers from a pool

**Consequences:**
- Decomposes parameter count from network depth
- Imposes inductive bias similar to RNNs
- Benefits optimization: accelerates learning of layers with similar functionality (redundant parameters $\rightarrow$ shared parameters)

**Parameters (M):**
- SWRN 28-2-1
- SWRN 28-4-2
- SWRN 28-10-2
- SWRN 28-16-2
- SWRN 28-18-2
- WRN 28-10
- WRN 40-4
- RNX 8x64
- RNX 16x64
- DN 40-12
- DN 100-24

**Test Error (%):**
- SWRN 28-2-1
- SWRN 28-4-2
- SWRN 28-10-2
- SWRN 28-16-2
- SWRN 28-18-2
- WRN 28-10
- WRN 40-4
- RNX 8x64
- RNX 16x64
- DN 40-12
- DN 100-24