DyRep: Bootstrapping Training with Dynamic Re-parameterization

Tao Huang[1,2] Shan You[1*] Bohan Zhang[3] Yuxuan Du[2] Fei Wang[4] Chen Qian[3] Chang Xu[2]

[1]SenseTime Research [2]School of Computer Science, Faculty of Engineering, The University of Sydney

[3]College of Information and Computer Sciences, University of Massachusetts Amherst

[4]University of Science and Technology of China

youshan@sensetime.com

Motivation

Structural re-parameterizaion (Rep):

1. **Train:** Transforms the original model into an augmented model.
2. **Inference:** Re-parameterizes the trained augmented model back to the original architecture.

Advantages:

1. Improves performance without increasing FLOPs or parameters.
2. Deploy-friendly: simple VGG-style architectures in inference.

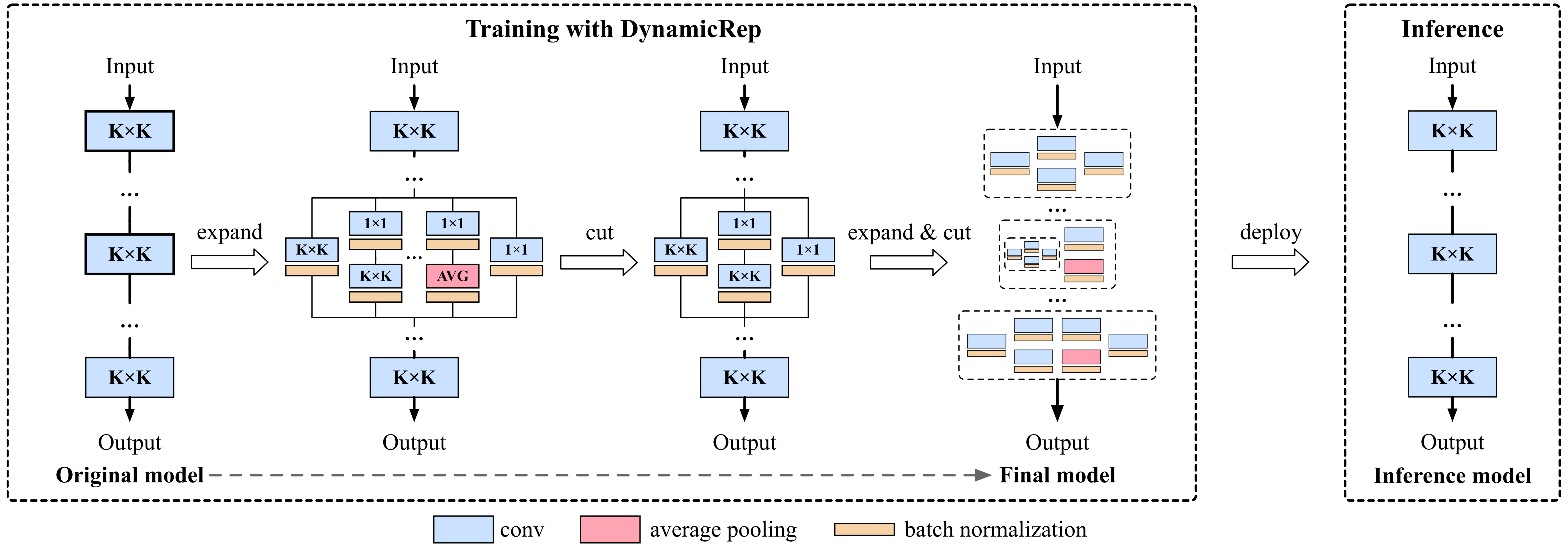
Limitations:

1. Expensive computation overhead: current methods simply Rep all operations into augmented ones.
2. Rep may introduce redundant operations, which slow down the training and even worsen the performance.

Intuition

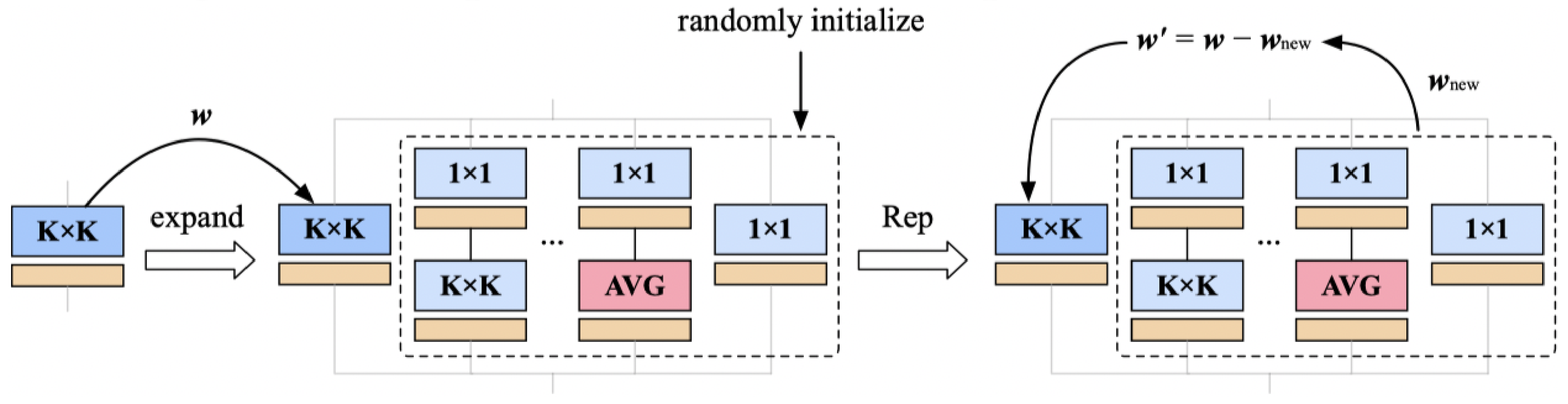
1. Only Rep those important operations in the model.
2. De-parameterize those redundant operations during training.

Dynamic Re-parameterization (DyRep)



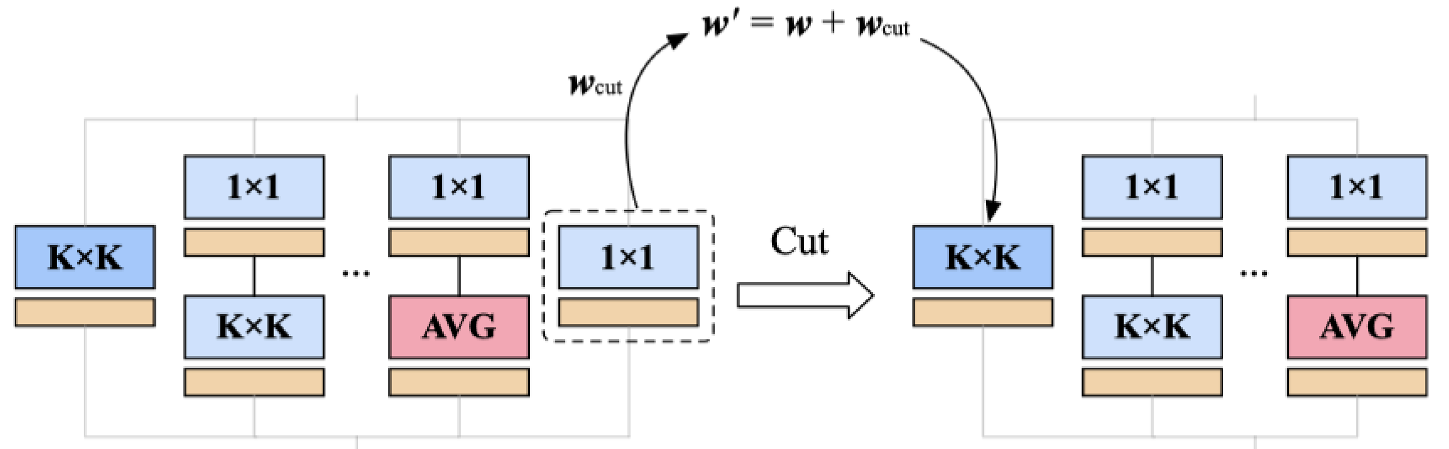
DyRep: Dynamically evolves network during training with Rep and Dep.

Training-aware re-parameterization (Rep):



1. Locate the operation contribute the most to the training loss using *synflow* saliency $\mathcal{S}_o(\boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial \theta_i} \odot \theta_i$.
2. Augment the located operation with additional branches, where the branches are randomly-initialized, and the weights original operation are computed using Rep.
3. Batch normalization layers are initialized and calibrated for stable training.

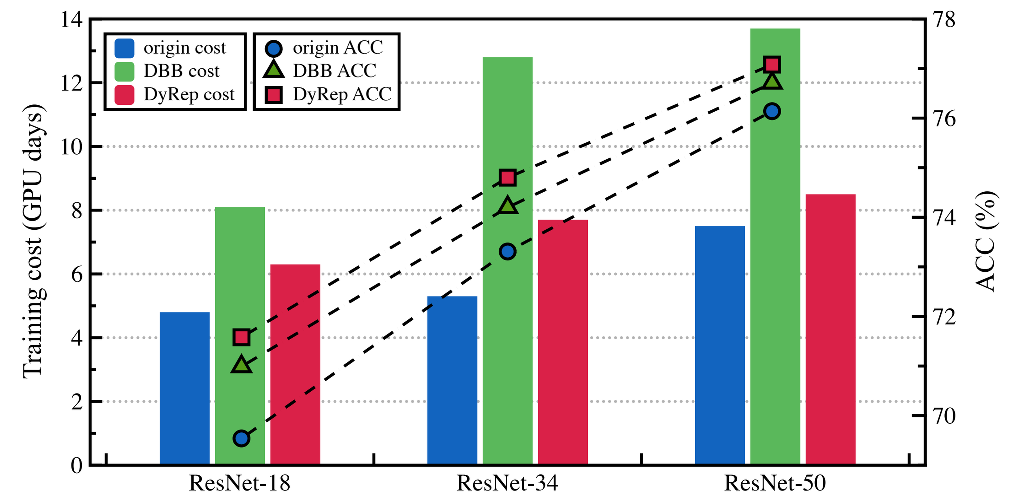
Training-aware de-parameterization (Dep):



1. Measure the importance of every branch using γ in BN: $s_j = \frac{1}{C}\sum_{k=1}^{C}|\gamma_k|$.
2. Locate the redundant operations with $s_j < \mathrm{Mean}(\{s_j\}_{j=1}^n)$ when the branches satisfying $\sqrt{\mathrm{Var}(\{s_j\}_{j=1}^n)}$.
3. Absorb the weights of removed operations into the original operation.

Experiments

Compare with DBB and plain networks:



Our DyRep obtains the highest accuracies yet has much smaller training cost compared to DBB.

Ablation on Rep and Dep:

Method	FLOPs (M)	Params (M)	ACC (%)
origin	313	15.0	94.68±0.08
Rep	658	29.3	95.03±0.15
Rep + Dep	597	29.3	**95.22±0.13**

Visualization of augmented operation: