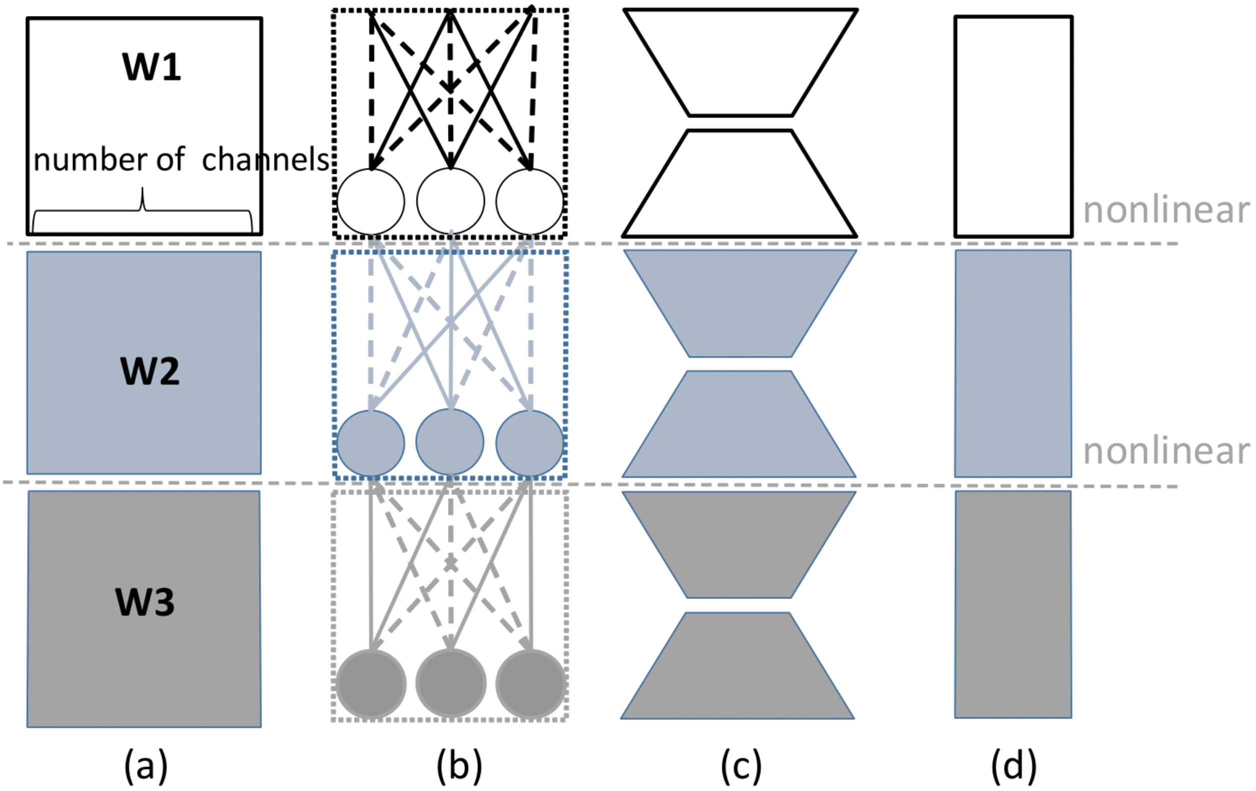
Browse > Methodology > Model Compression



Model Compression

Edit

44 papers with code · Methodology

Leaderboards

Add a Result

You can find evaluation results in the subtasks. You can also submitting evaluation metrics for this task.

Subtasks

Neural Network Compression

2 leaderboards

21 papers with code

Greatest papers with code

GreatestLatestWithout code

The State of Sparsity in Deep Neural Networks

25 Feb 2019 • google-research/google-research • TensorFlow

We rigorously evaluate three state-of-the-art techniques for inducing sparsity in deep neural networks on two large-scale learning tasks: Transformer trained on WMT 2014 English-to-German, and ResNet-50 trained on ImageNet.

MODEL COMPRESSION

★ 6,725

Paper

Code

SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size

24 Feb 2016 • pytorch/vision • PyTorch

(2) Smaller DNNs require less bandwidth to export a new model from the cloud to an autonomous car.

MODEL COMPRESSION

★ 5,429

Paper

Code

Model compression via distillation and quantization

ICLR 2018 • NervanaSystems/distiller • PyTorch

Deep neural networks (DNNs) continue to make significant advances, solving tasks from image classification to translation or reinforcement learning.

MODEL COMPRESSION QUANTIZATION

★ 2,591

Paper

Code

AMC: AutoML for Model Compression and Acceleration on Mobile Devices

ECCV 2018 • NervanaSystems/distiller • PyTorch

Model compression is a critical technique to efficiently deploy neural network models on mobile devices which have limited computation resources and tight power budgets.

MODEL COMPRESSION NEURAL ARCHITECTURE SEARCH

★ 2,591

Paper

Code

Global Sparse Momentum SGD for Pruning Very Deep Neural Networks

NeurIPS 2019 • ShawnDing1994/ACNet • PyTorch

Deep Neural Network (DNN) is powerful but computationally expensive and memory intensive, thus impeding its practical usage on resource-constrained front-end devices.

MODEL COMPRESSION

★ 406

Paper

Code

Contrastive Representation Distillation

23 Oct 2019 • HobbitLong/RepDistiller • PyTorch

We demonstrate that this objective ignores important structural knowledge of the teacher network.

MODEL COMPRESSION TRANSFER LEARNING

★ 390

Paper

Code

Data-Free Knowledge Distillation for Deep Neural Networks

19 Oct 2017 • huawei-noah/DAFL • PyTorch

Recent advances in model compression have provided procedures for compressing large neural networks to a fraction of their original size while retaining most if not all of their accuracy.

MODEL COMPRESSION

★ 270

Paper

Code

Discrimination-aware Channel Pruning for Deep Neural Networks

NeurIPS 2018 • SCUT-AILab/DCP • PyTorch

Channel pruning is one of the predominant approaches for deep model compression.

MODEL COMPRESSION

★ 130

Paper

Code

A Programmable Approach to Model Compression

6 Nov 2019 • NVlabs/condensa • PyTorch

However, while the results are desirable, finding the best compression strategy for a given neural network, target platform, and optimization objective often requires extensive experimentation.

IMAGE CLASSIFICATION LANGUAGE MODELLING MODEL COMPRESSION QUANTIZATION

90

Paper

Code

MicroExpNet: An Extremely Small and Fast Model For Expression Recognition From Face Images

19 Nov 2017 • cuguilke/microexpnet • TensorFlow

On the other hand, KD is proved to be useful for model compression for the FER problem, and we discovered that its effects gets more and more significant with the decreasing model size.

FACIAL EXPRESSION RECOGNITION MODEL COMPRESSION

85

Paper

Code

Contact us on: hello@paperswithcode.com. Papers With Code is a free resource supported by Atlas ML.

Terms Privacy Cookies policy