上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Institute of Media,
Information, and Network

Efficient Decoder for Learned Image Compression via Structured Pruning

Liewen Liao, Shaohui Li, Jixiang Luo,
Wenrui Dai, Chenglin Li, Junni Zou,
and Hongkai Xiong

liaolw@sjtu.edu.cn

Department of Electronic Engineering

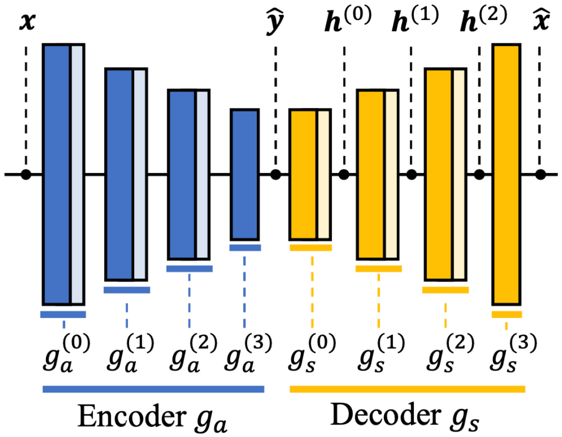
Shanghai Jiao Tong University

Contents

- Background & Motivations
- Methodology
- Experimental Results
- Outlook

Background & Motivations

- Symmetric architecture in Learned Image Compression:



- Varying computational resources in different decoding devices:



- Customize models to accommodate different decoding platforms;

Methodology

- Effectiveness evaluation via activation range estimation:

Algorithm 1 Network Pruning on Learned Decoder

Input: A pre-trained learned compression model with encoder g_a and decoder g_s, the initial value for input $\boldsymbol{x}_0$, the number of iterations for gradient ascent N, the number of pruned channels on each layer s_i.

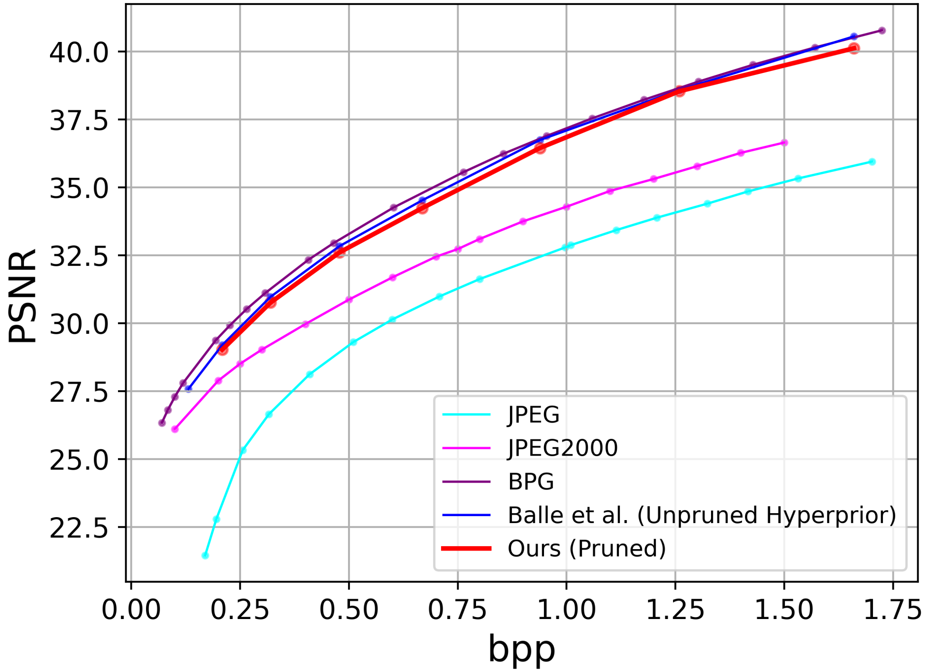
Output: A pruned decoder $\tilde{g}_s$.

for $i \leftarrow 2$ to 0 **do** ▷ Each layer in the decoder.
  for $j \leftarrow 1$ to q **do** ▷ Each channel of i-th layer.
    $n \leftarrow 0, \boldsymbol{x} \leftarrow \boldsymbol{x}_0$
    while $n \leq N$ and not converged **do** ▷ Loop for gradient ascent.
      $\boldsymbol{x} \leftarrow \boldsymbol{x} + \eta \cdot (\partial \dot{\boldsymbol{h}}_j^{(i)} / \partial \boldsymbol{x})$
      $n \leftarrow n + 1$
    end while
    while $n \leq N$ and not converged **do** ▷ Loop for gradient descent.
      $\boldsymbol{x} \leftarrow \boldsymbol{x} - \eta \cdot (\partial \ddot{\boldsymbol{h}}_j^{(i)} / \partial \boldsymbol{x})$
      $n \leftarrow n + 1$
    end while
    $e_j^{(i)} \leftarrow \dot{\boldsymbol{h}}_j^{(i)} - \ddot{\boldsymbol{h}}_j^{(i)}$ ▷ Record the effectiveness.
  end for
  for $k \leftarrow 1$ to s_i **do**
    Remove the j^*-th ineffective channel.
  end for
end for
Fine-tune the pruned model and return $\tilde{g}_s$.

Experimental Results

- Pruning performance:

Pruning Ratio r	Unpruned	0.20	0.33	0.40	0.55
Model Size (MB)	1.49	1.20	0.99	0.88	0.70
PSNR (dB)	32.84	32.75	32.66	32.61	32.43
PSNR decay (dB/MB)	-	-0.31	-0.32	-0.37	-0.52
Inference Time (s)	0.71	0.60	0.55	0.53	0.48
FLOPS (M)	3618	2656	1899	1533	1141



Outlook

- Pruning methods for image compression:
  - This paper focus on pruning pre-trained model to produce light-weighted models;
  - Design one-stage methods perform pruning and training simultaneously;
  - Novel light-weighted architecture;

- General network acceleration methods:
  - Quantization;
  - Distillation;
  - …

Institute of Media,
Information, and Network

Thank You!

Liewen Liao, Shaohui Li, Jixiang Luo,
Wenrui Dai, Chenglin Li, Junni Zou,
and Hongkai Xiong

liaolw@sjtu.edu.cn

Department of Electronic Engineering

Shanghai Jiao Tong University