Learning Flows By Parts

Manush Bhatt, David I. Inouye
Purdue University

PURDUE UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Definition

Can state-of-the-art normalizing flow models be trained without full end-to-end backpropagation but still attain reasonable performance?

Our Answer

We optimize flows in gradient-isolated parts via truncated objective. This prevents end-to-end backpropagation between flow parts.

Potential Benefits

- Reduced computational burden
- Reaches comparable accuracy in shorter time
- Opens possibility of training across multiple devices

Local/Truncated Objective Function

The probabilistic model of normalizing flows using a known prior distribution $p_z(z)$ over the latent variable z and an invertible function f follows as:

$$\log p_x(x) = \log p_z(f(x)) + \log\left|\det \frac{\partial f(x)}{\partial x}\right|,$$

where $\frac{\partial f(x)}{\partial x}$ is the Jacobian of f. The function f is composed of invertible functions, i.e. $f = f_L \circ f_{L-1} \circ f_{L-2} \circ \cdots \circ f_1$ where L denotes the number of layers or modules.

The Maximum Likelihood Estimation can be written as:

$$\min_{f_1, f_2, \ldots, f_L} -\log p_z(f(x)) - \sum_{l=1}^{L} \log\left|\det \frac{\partial f_l(z^{(l-1)})}{\partial z^{(l-1)}}\right|,$$

where $z^{(l-1)} = f_{l-1} \circ \cdots \circ f_1(x)$.

We can **locally optimize** the k^{th} layer by dropping the constant terms with respect to f_k (the log det terms) and evaluating the prior term only based on the output of f_k (ignoring future layers):

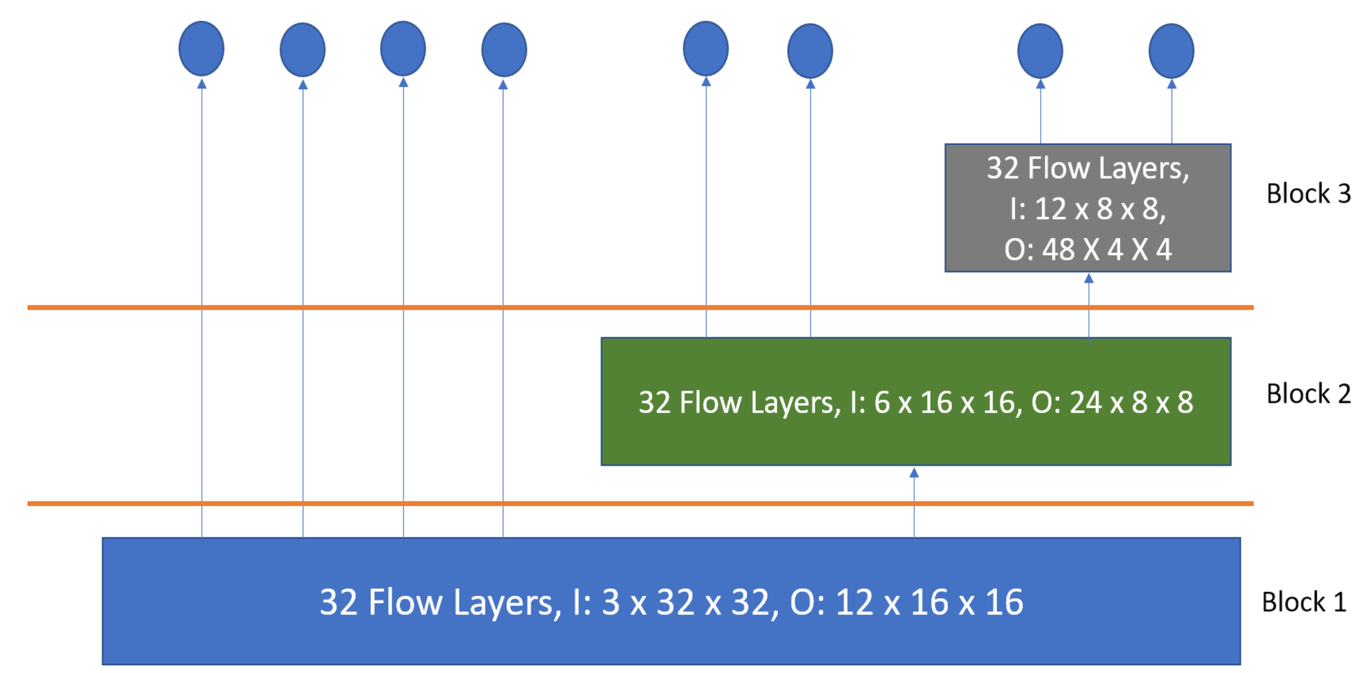
$$\min_{f_k} -\log p_z\left(f^{(1:k)}(x)\right) - \log\left|\det \frac{\partial f_k(z^{(k-1)})}{\partial z^{(k-1)}}\right|,$$

where $f^{(1:k)} = f_k \circ f_{k-1} \circ \cdots \circ f_1$. Each f_k term is only dependent on previous layers through latent representation $z^{(k-1)}$ and this can be viewed as **truncated objective** function.

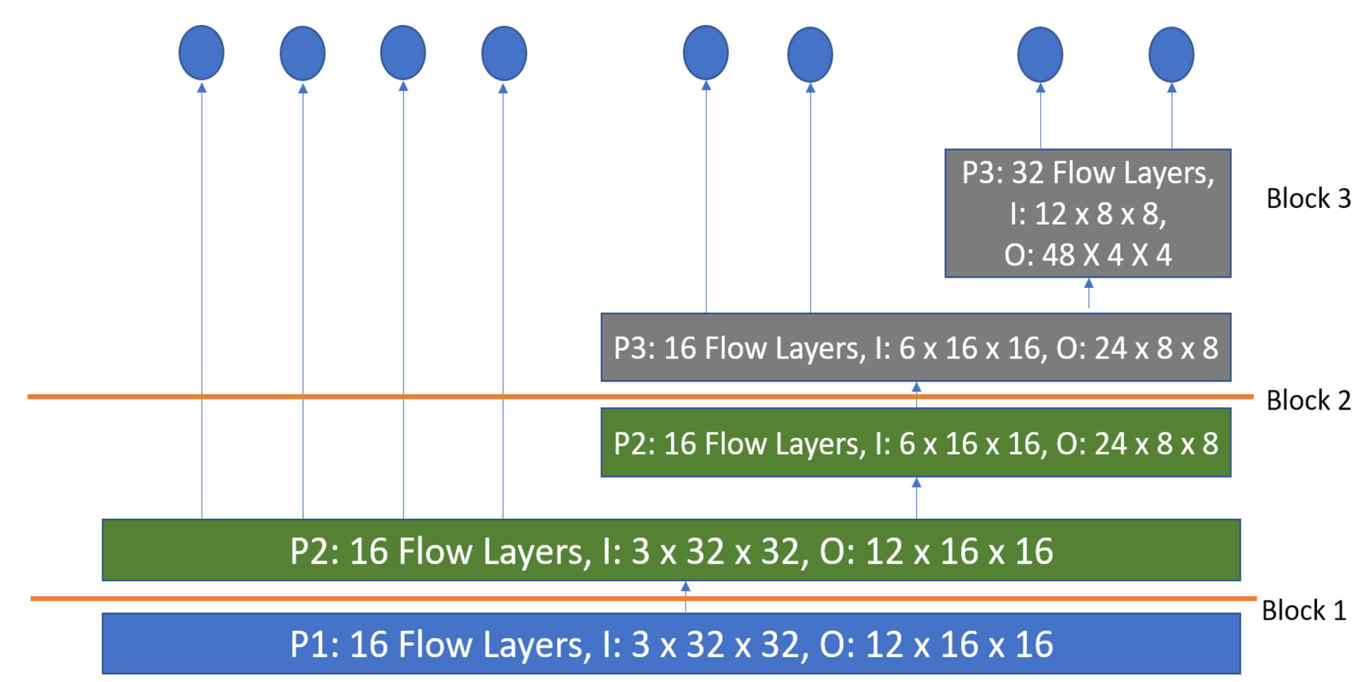
Glow Model [1]

- Contains 3 high-level hierarchical blocks with squeeze op.
- Between blocks half of the channels continue through flow.
- Each block has 32 flow layers (Actnorm, Invertible 1x1 Convolution and Affine Coupling).
- We split model into **gradient-isolated parts** using two schemes: Split By Blocks or Split Across Blocks.

Split **By** Blocks



Split **Across** Blocks



Training Mechanism

Sequential Training (unstable)

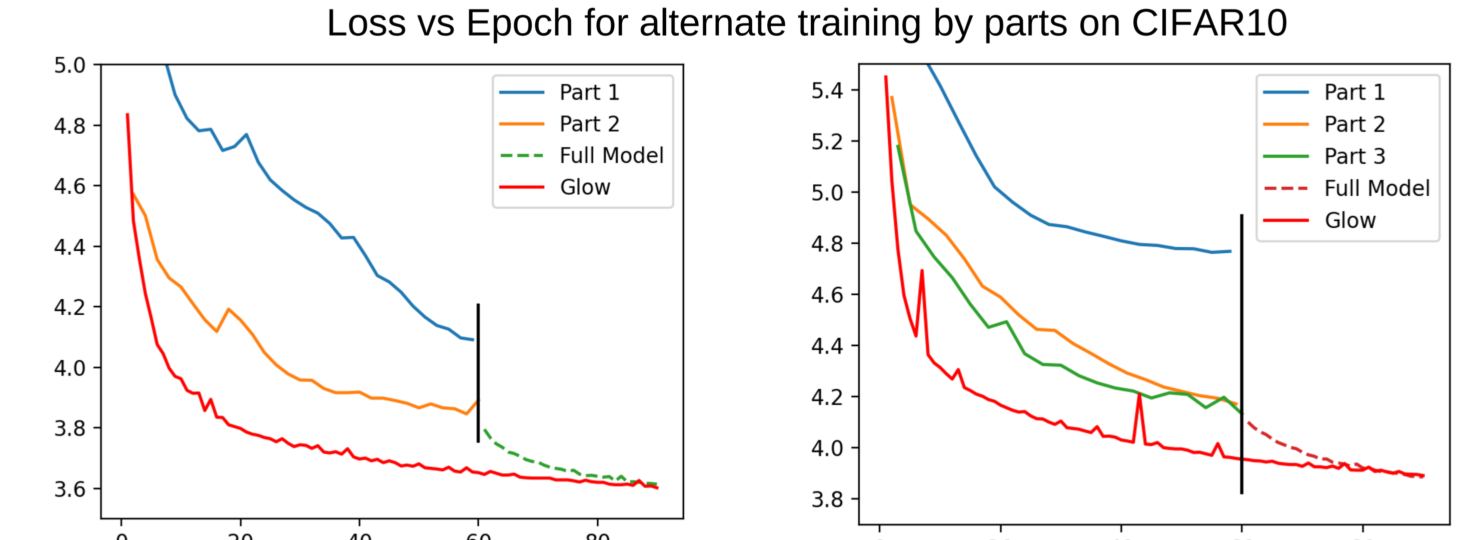
Training each Glow part greedily where each part is trained for a fixed number of epochs followed by the next. Our experiments show that this leads to instability when optimizing the second part after the first part has reasonably converged.

Alternate Training (stable)

We propose an alternate training mechanism where each part is alternatively trained using weights of the previous part.

Results

Loss vs Epoch for alternate training by parts on CIFAR10



(a) GLOW with 2-part split-across-blocks

(b) GLOW with 3-part split-across-blocks with a lower learning rate

We achieve results that **match full end-to-end training** by alternate training of Glow parts for the initial 60 epochs (gradients do not flow between parts) followed by 30 epochs of end-to-end global optimization.

Approach	No. of Parts	Parts	LR	BPD	Time (mins)
Glow	1	1,2,3	1e-4	3.60	1026
Ours (Across)	2	$1_{1/2}$ -> $1_{1/2}$,2,3	1e-4	3.61	**720**
Glow	1	1,2,3	1e-5	3.89	1020
Ours (Across)	3	$1_{1/2}$ -> $1_{1/2}$,$2_{1/2}$ -> $2_{1/2}$,3	1e-5	3.89	**647**

Our approach **saves at least 30% of time** compared to full end-to-end backpropagation as shown above. Additionally, we emphasize that our approach also **reduces the backward communication** needed between the part optimizations, which could be useful in a distributed environment.

Next Steps

- Can we completely remove the last few epochs of end-to-end backpropagation and still achieve similar results?
- Is there an optimal way to find the learning rates for each part?
- Can we stabilize the training via other alternatives?

References

D. P. Kingma and P. Dhariwal (2018). "Glow: Generative flow with invertible 1x1 convolutions" In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 10215–10224. Curran Associates, Inc

S. Löwe, P. O'Connor, and B. Veeling, "Putting an end to end-to-end: Gradient-isolated learning of representations" In: Advances in Neural Information Processing Systems, pages 3039–3051, 2019