AC-GC: LOSSY ACTIVATION COMPRESSION WITH GUARANTEED CONVERGENCE

R. David Evans, Tor M. Aamodt

Neural Information Processing Systems
(NeurIPS), December 7, 2021
Poster Spot A2, 1630 PST – 1800 PST

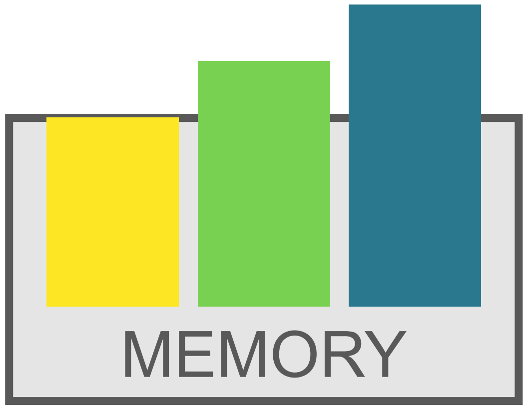
THE UNIVERSITY OF BRITISH COLUMBIA
Electrical and Computer Engineering

NSERC CRSNG
COHESA

AC-GC: IN A NUTSHELL

Neural network training
cares about

Lossy Compression
decreases
memory footprint



How to set
Compression Rate

image source: nicePNG

MOTIVATION

KEY CONTRIBUTIONS

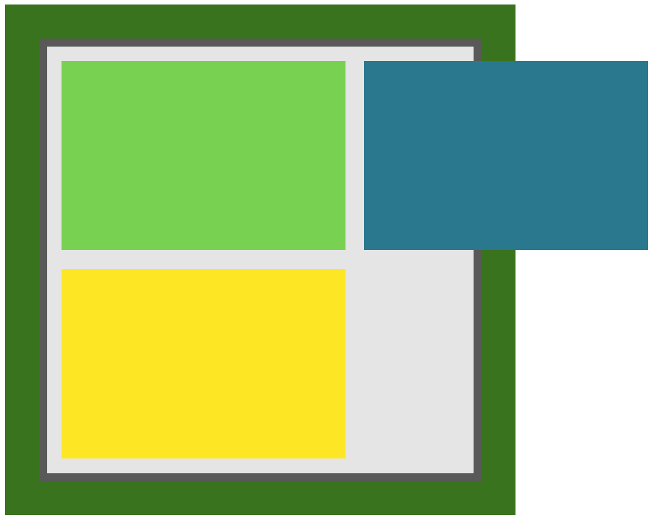
BOUNDING CONVERGENCE

SIMPLIFYING THE PROBLEM

COMPRESSION

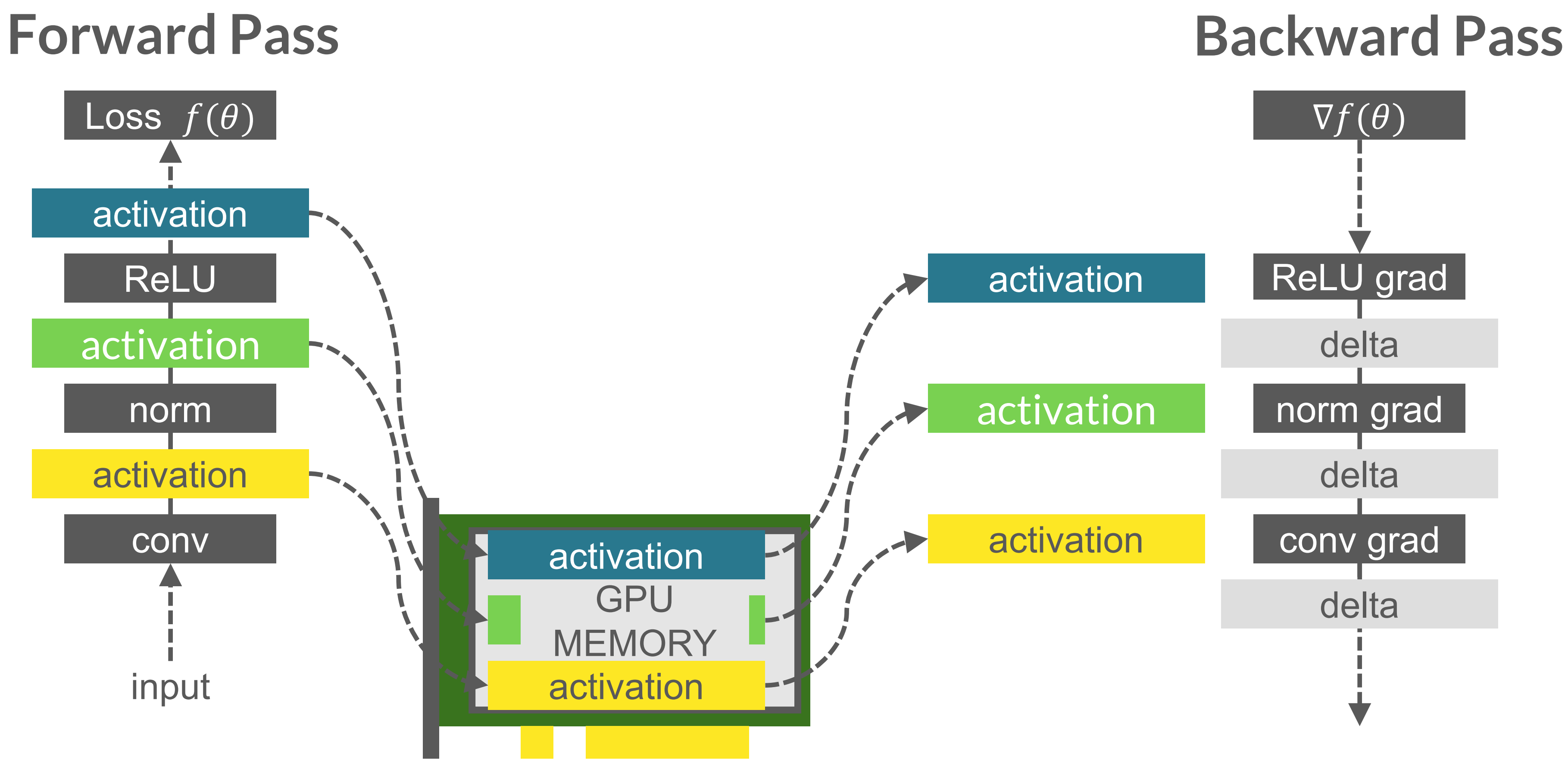
RESULTS

MOTIVATION



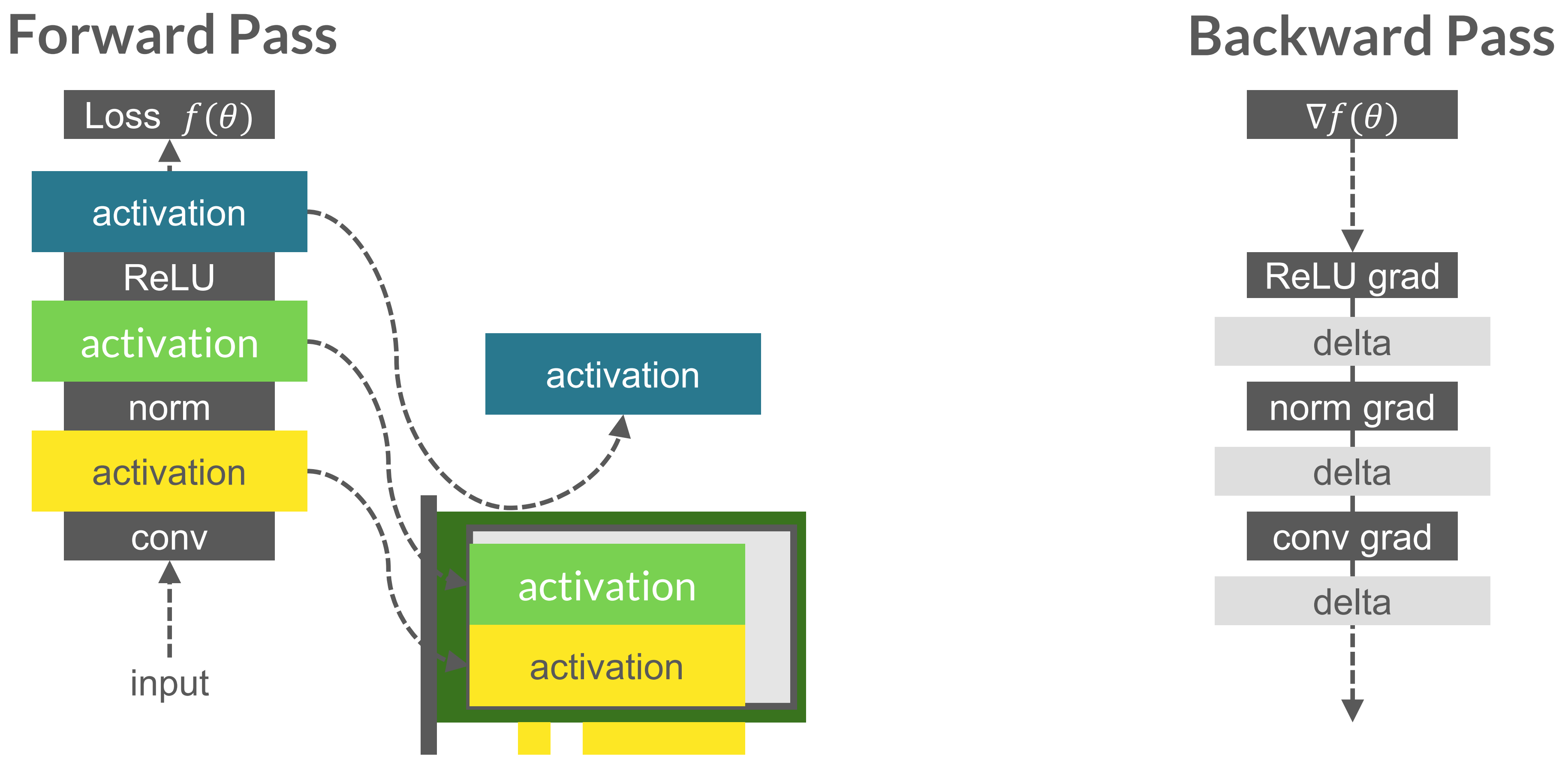
ACTIVATIONS DURING TRAINING

Activations are memoized layer outputs from the forward pass

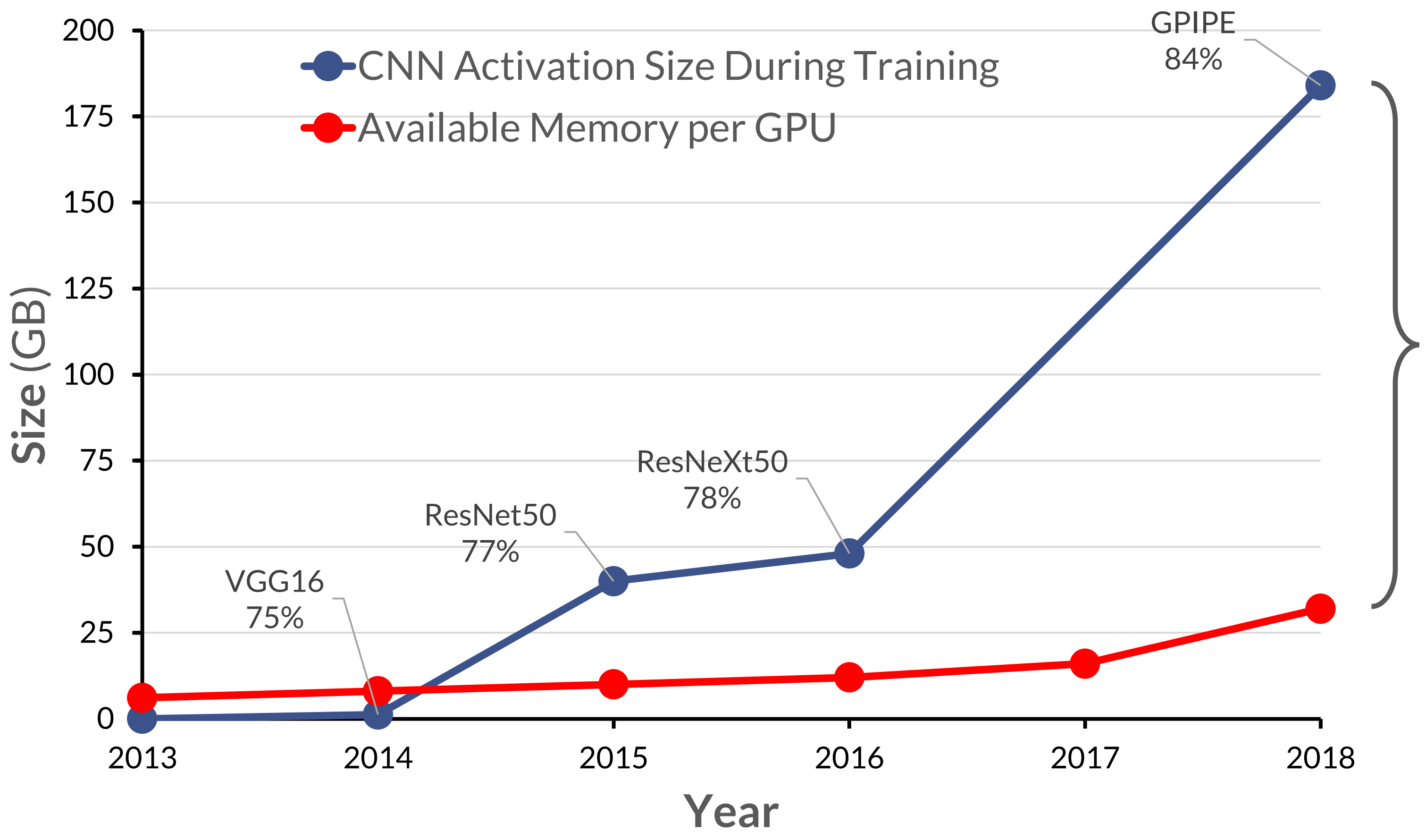


LOSSY ACTIVATION COMPRESSION

Larger networks and larger batch sizes mean more activations



THE ACCELERATOR MEMORY WALL



Due to **physical limitations**, e.g. bandwidth and memory technology

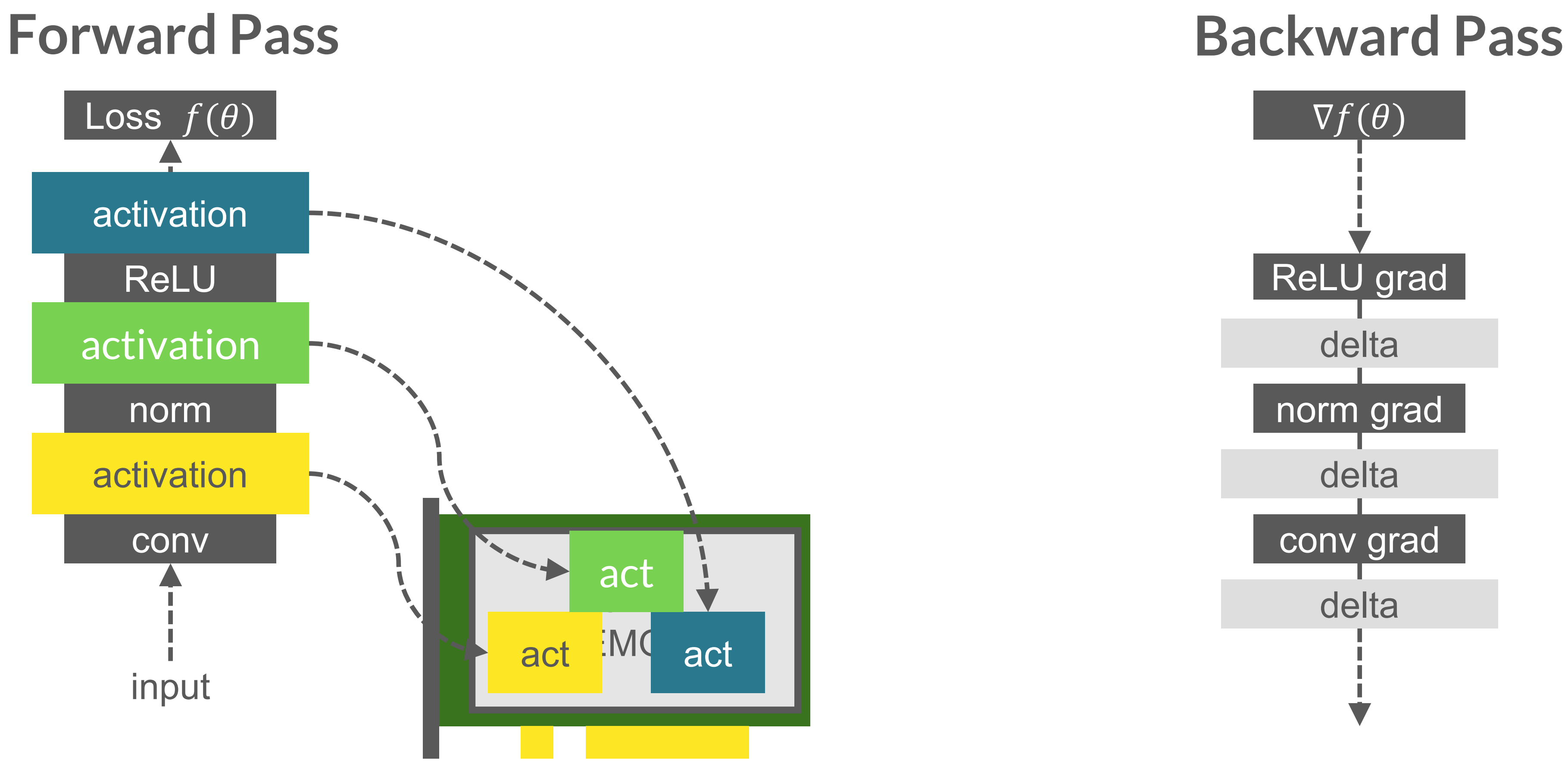
DECREASING ACTIVATION MEMORY

Many orthogonal approaches

Approach	Training Throughput	Memory Reduction	Accuracy after Training
	Relative to normal training		
Scheduling e.g. GIST (A Jain, ISCA 2018)	✗	✓	
Recalculation e.g. Sublinear Nets (T. Chen, ArXiv 2016)		✓✓	
Offloading e.g. vDNN (M. Rhu, MICRO 2016)	✗✗✗✗	✓✓✓✓	
Lossless Compression e.g. CDMA (M. Rhu, HPCA 2018)	✗✗	✓✓	
Lossy Compression e.g. ACTNN (J. Chen, ICML 2021)	✗✗	✓✓✓✓✓	✗

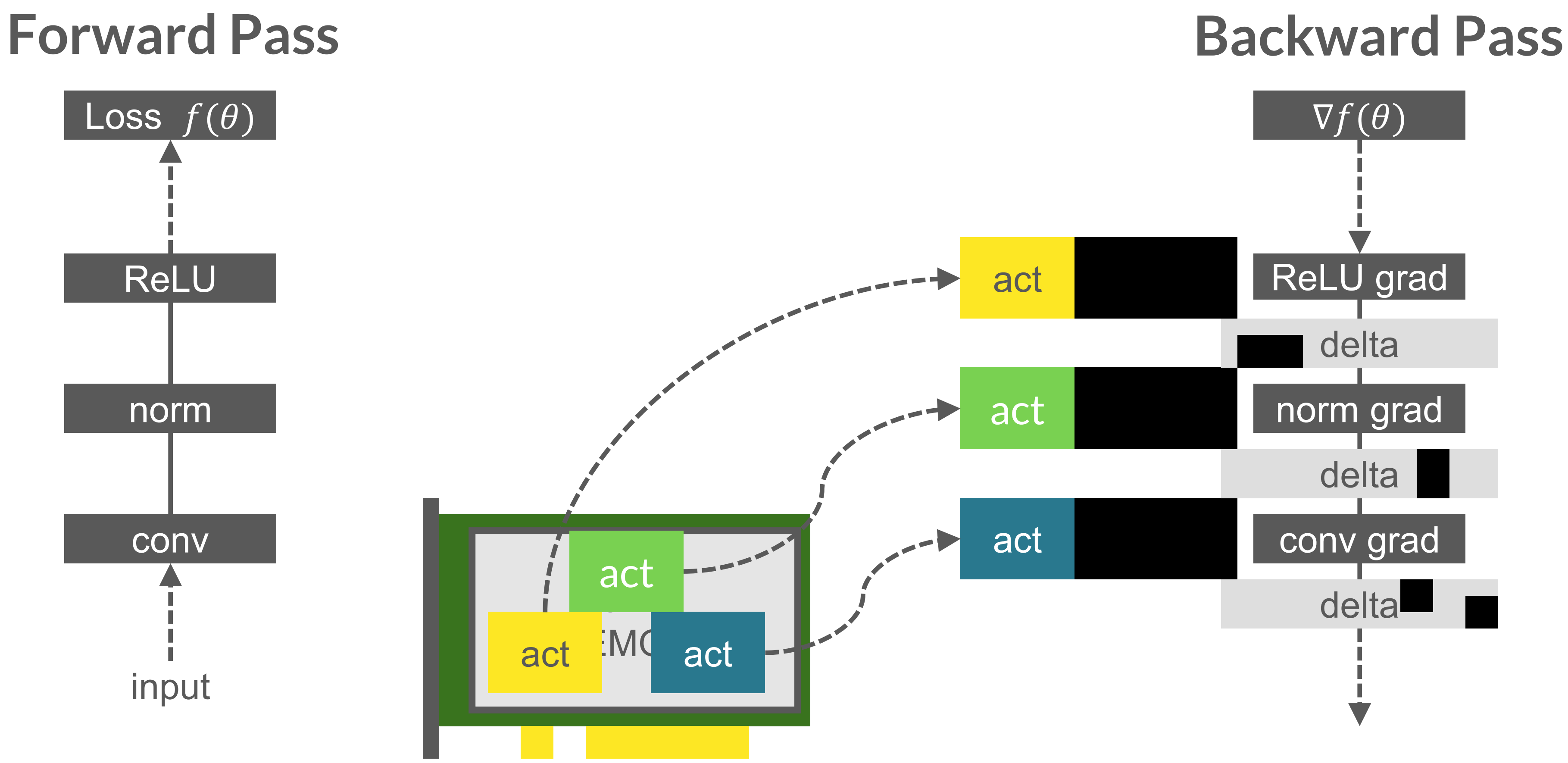
LOSSY ACTIVATION COMPRESSION

Discard some data, and compress



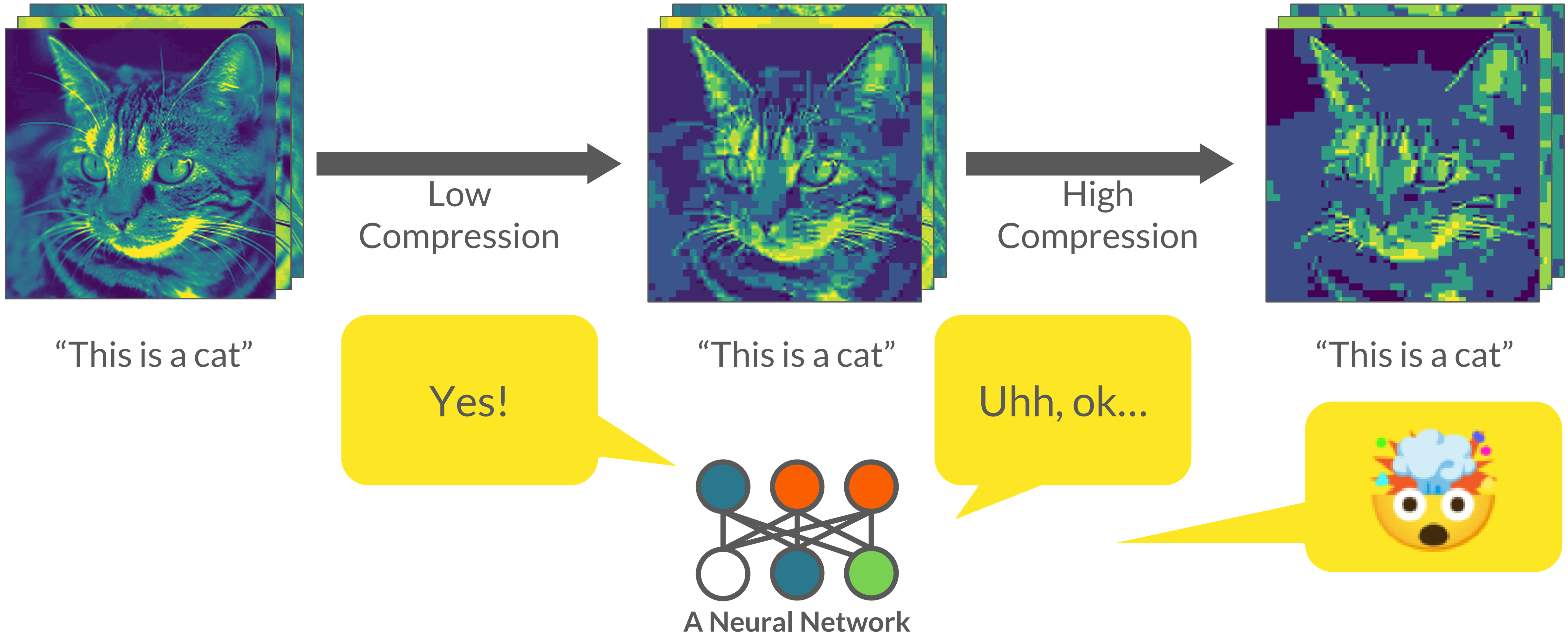
LOSSY ACTIVATION COMPRESSION

Discard some data, and compress

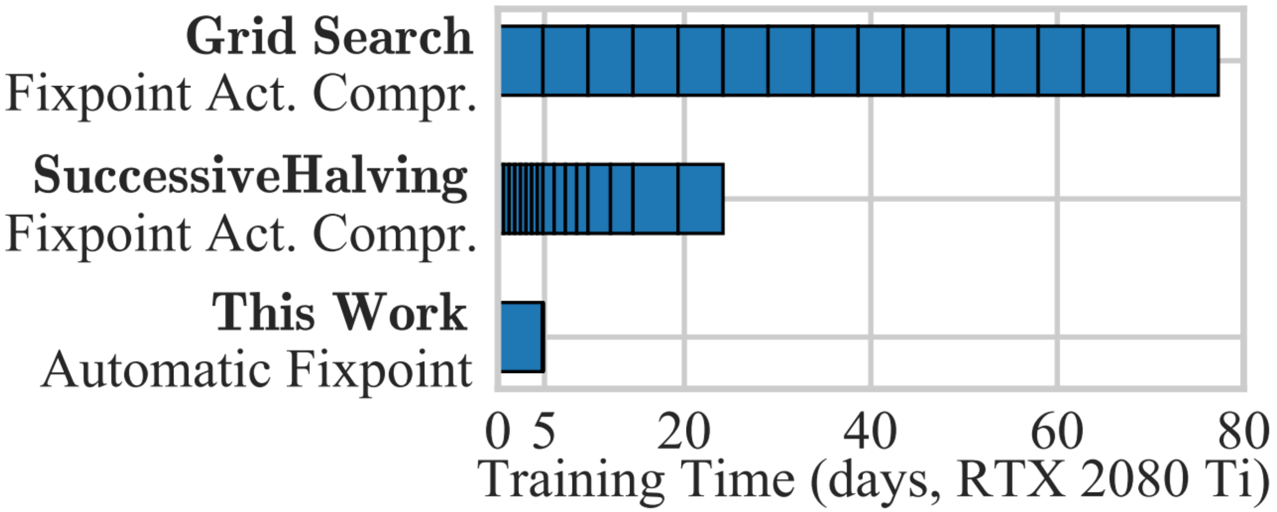


DRAWBACKS OF LOSSY COMPRESSION

How to avoid explosions during training?



TUNING THE COMPRESSION RATE IS HARD



Compression rate search cost. Each box indicates a different compression rate (1- to 16-bit fixpoint)

KEY CONTRIBUTIONS

CONSTRAINED OPTIMIZATION FOR LOSSY ACTIVATIONS

Maximize Act. Compression

$$\underset{\Delta X}{\operatorname{argmax}}\, B(\Delta X)$$

$s.t.$

Persevering Convergence

Goal: An efficient way to get compression rate from convergence

CONSTRAINED OPTIMIZATION FOR LOSSY ACTIVATIONS

Maximize Act. Compression

$$\underset{\Delta X}{\operatorname{argmax}}\, B(\Delta X)$$

$s.t.$

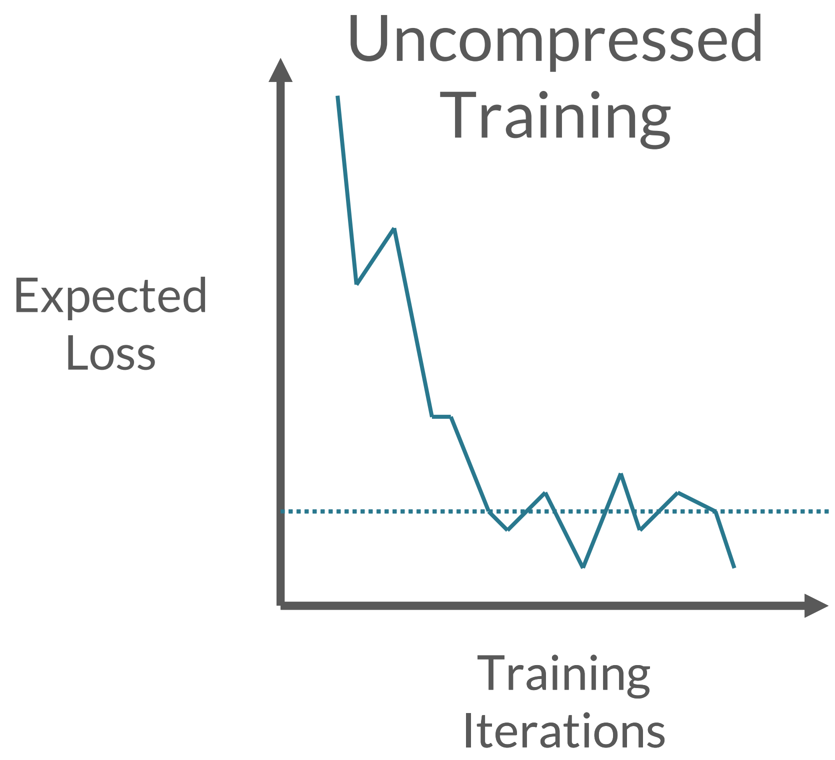
Persevering Convergence ?

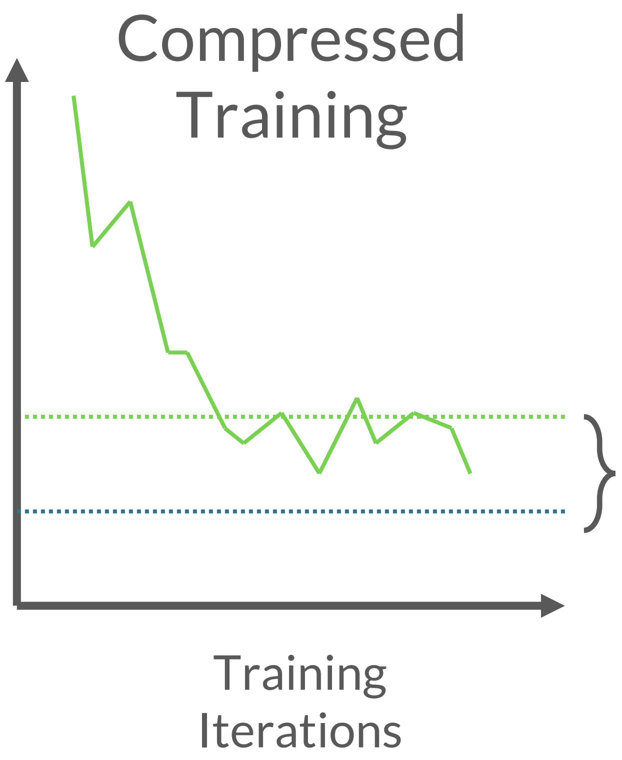
Activation **error**

Some measure of **compression**

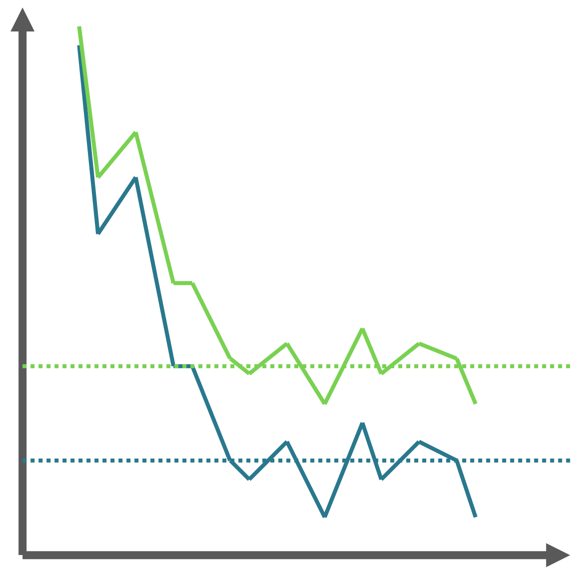
KEY INSIGHT: ALLOW LOSS TO INCREASE

Takes advantage of **SGD convergence** behaviour





BOUNDING CONVERGENCE



STOCHASTIC GRADIENT DESCENT NOTATION

Overall loss

Loss of example n

$$\mathcal{L}(\theta) = \sum_n f_n(\theta)$$

Parameters

Loss is a finite sum

Solve using S.G.D.

Gradient w.r.t. θ

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \nabla_\theta f_{n_t}(\theta)$$

Loss of a randomly selected training example

SGD THEORETICAL CONVERGENCE RATES

Many results on SGD convergence rates, we use:

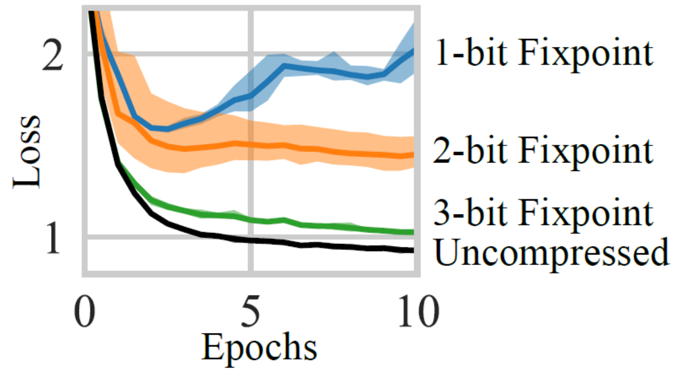
$$\boxed{\mathbb{E}\left[\mathcal{L}\left(\theta^{(t)}\right) - \mathcal{L}\left(\theta^{(*)}\right)\right]} \leq \boxed{(1 - C_1\alpha)^t}\left(\mathcal{L}\left(\theta^{(0)}\right) - \mathcal{L}\left(\theta^{(*)}\right)\right) + C_2\alpha V^2$$

C_1, C_2: Constants
V^2: Gradient Variance

[1] H Karimi, J Nutini, M Schmidt, ***"Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition,"*** ECML PKDD 2016

ALLOWING ERRORS FOR COMPRESSION

CIFAR10/ResNet50



Uncompressed Training

Expected Loss

$\mathbb{E}[\mathcal{L}(\theta^{(t)}) - \mathcal{L}(\theta^{(*)})]$

Training Iterations

$\propto V^2$

Grad. Variance

Compressed Training

Training Iterations

$\propto (1 + e^2)V^2$

Compr. Grad. Variance

GRADIENT VARIANCE

Compressed training can be viewed as SGD with increased variance

From Karimi et. al:

Gradient **"variance"** is bounded:

$$\mathbb{E}\left[\left\|\nabla_\theta f_{n_t}(\theta)\right\|^2\right] \leq V^2$$

Under compression:

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\theta f_{n_t}(\theta)\right\|^2\right] \leq (1+e^2)V^2$$

CONSTRAINED OPTIMIZATION FOR LOSSY ACTIVATIONS

Gradient variance is bounded

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\theta f\right\|^2\right] \leq (1+e^2)V^2$$

Maximize Act. Compression

$$\underset{\Delta X}{\operatorname{argmax}}\, B(\Delta X)$$

$s.t.$

Persevering Convergence

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\theta f_{n_t}(\theta)\right\|^2\right] \leq (1+e^2)V^2$$

ERROR-VARIANCE RELATIONSHIP

Using an additive gradient error:

$$\widehat{\nabla}_\theta f \equiv \nabla_\theta f + \Delta\nabla_\theta f$$

The constraint on variance: $\xrightarrow{\left(\substack{\text{some math with} \\ \text{expectations and norms}}\right)}$ Is satisfied by:

$$\mathbb{E}[\|\nabla_\theta f + \Delta\nabla_\theta f\|^2] \leq (1 + e^2)V^2$$

$$\|\Delta\nabla_\theta f\|^2 \leq e^2V^2$$

CONSTRAINED OPTIMIZATION FOR LOSSY ACTIVATIONS

Maximize Act. Compression

$$\underset{\Delta X}{\operatorname{argmax}} B(\Delta X)$$

$s.t.$

Persevering Convergence

Gradient variance is bounded

$$\mathbb{E}\left[\left\|\widehat{\nabla}_\theta f\right\|^2\right] \leq (1+e^2)V^2$$

Gradient error is bounded

$$\|\Delta\nabla_\theta f\|^2 \leq e^2 V^2$$

SIMPLIFYING THE PROBLEM

BOUNDING FUNCTION

Gradient error is related to activation error

Now we just need to calculate it for a layer, e.g., convolution…

$$\|\Delta\nabla_\theta f\|^2 = \cancel{\sum_{k,c,r,s}^{K,C,R,S} \left(\sum_{n,h,w}^{N,H,W} \Delta x_{n,c,h+r,w+s} \frac{\partial f}{\partial y_{nkhw}} \right)^2} \leq e^2 V^2$$

which is not very useful… the solution is not closed-form. Instead use:

$$\|\Delta\nabla_\theta f\|^2 \leq D(\Delta X) \leq e^2 V^2 \longrightarrow$$

Where we find $D(\Delta X)$ that is as close as possible to the error norm.

CONSTRAINED OPTIMIZATION FOR LOSSY ACTIVATIONS

Gradient variance is bounded

$$\mathbb{E}\left[\left\|\hat{\nabla}_\theta f\right\|^2\right] \leq (1+e^2)V^2$$

Maximize Act. Compression

$$\underset{\Delta X}{\operatorname{argmax}}\, B(\Delta X)$$

$s.t.$

Persevering Convergence

$$D(\Delta X) \leq e^2 V^2$$

Gradient error is bounded

$$\|\Delta \nabla_\theta f\|^2 \leq D(\Delta X)$$

APPROXIMATING NORMS

Two Issues with calculating activation errors this way:

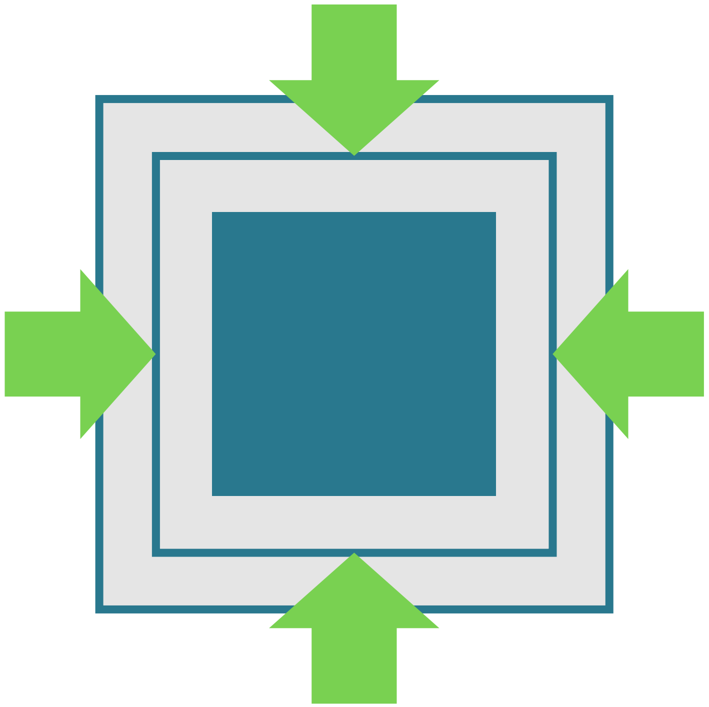
Calculating the norms are expensive

Calculate norms only every 100 iterations

Gradient information is not available in the forward pass

Estimate the norms using a rolling mean

COMPRESSION



COMPRESSION METRIC

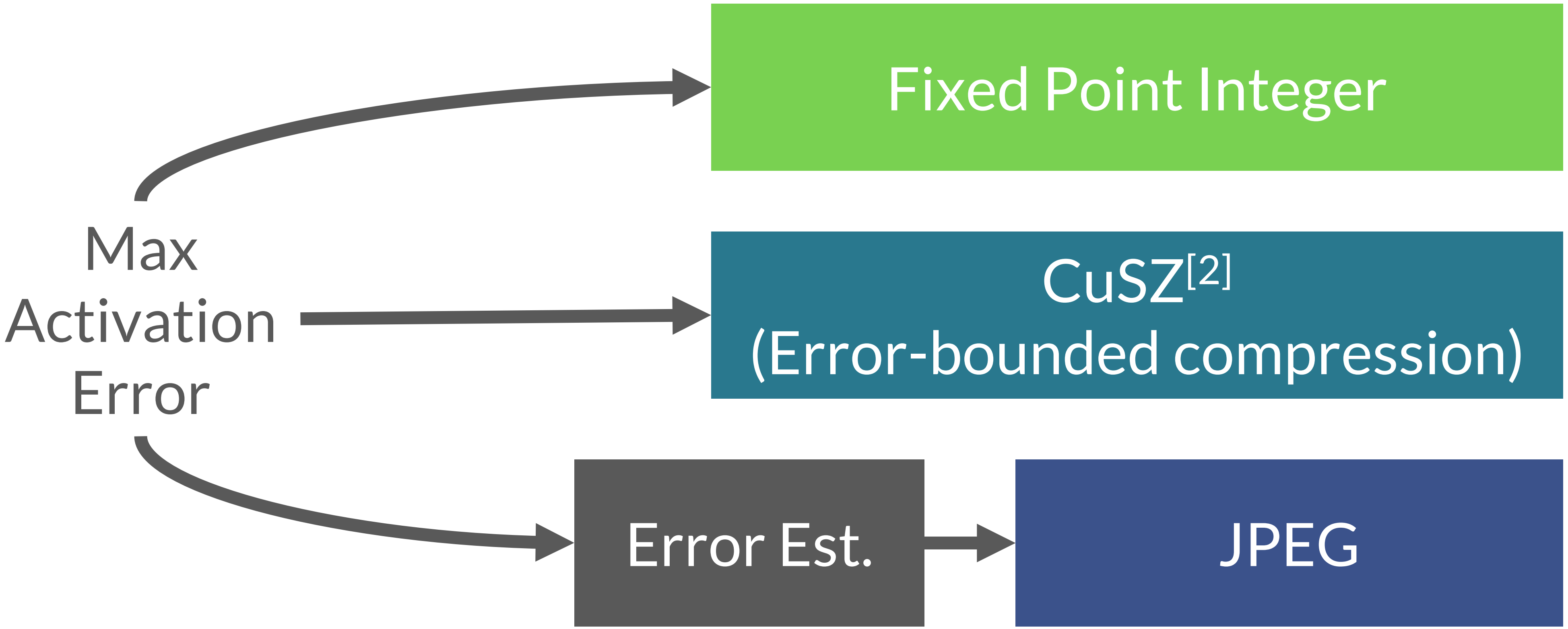
Choose B given that most methods use precision reduction

$$B(\Delta X) \propto Number\ of\ bits\ removed$$

$$B(\Delta X) \equiv \sum_i \log|\Delta x_i|$$

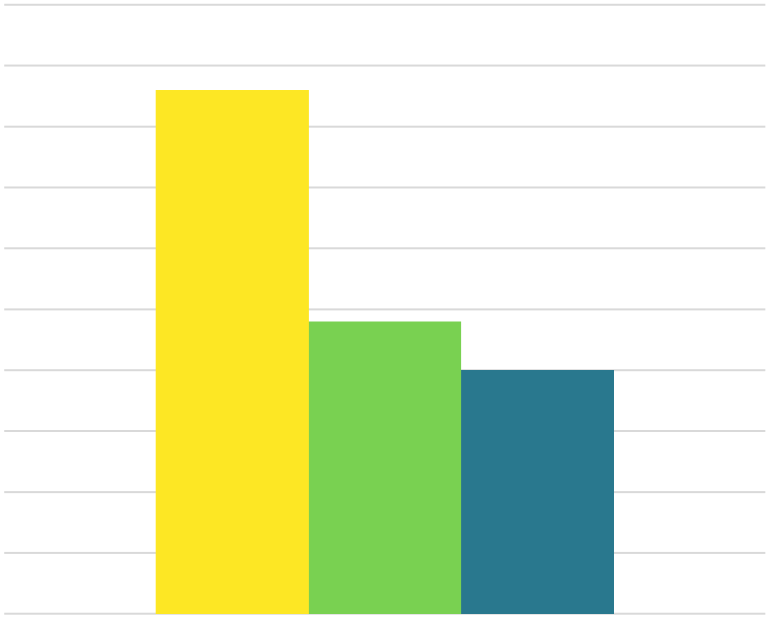
COMPRESSION METHODS

Case studies on activation error-bounded compression methods



[2] S. Jin, G. Li, S. L. Song, D. Tao, ***"A Novel Memory-Efficient Deep Learning Training Framework via Error-Bounded Lossy Compression"***, *in ArXiv 2020*

RESULTS



ERROR BOUNDS ARE DERIVED PER-LAYER

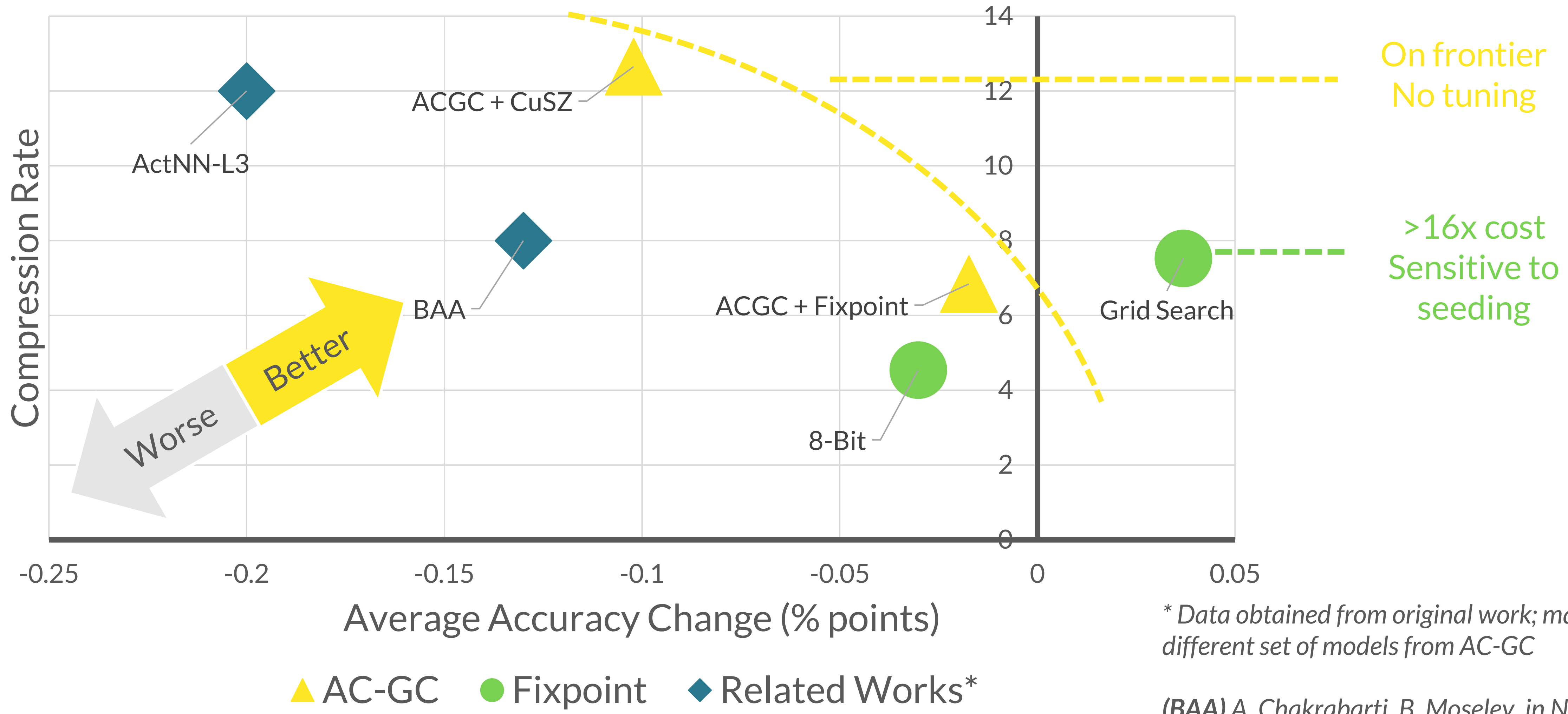
linear conv norm ReLU sum dropout

$$(\Delta x_{nchw})^2 \leq \boxed{e^2 V^2} \cdot \frac{1}{RSNCHW \|\nabla_Y f\|^2}$$

Filter size

Activation dimensions

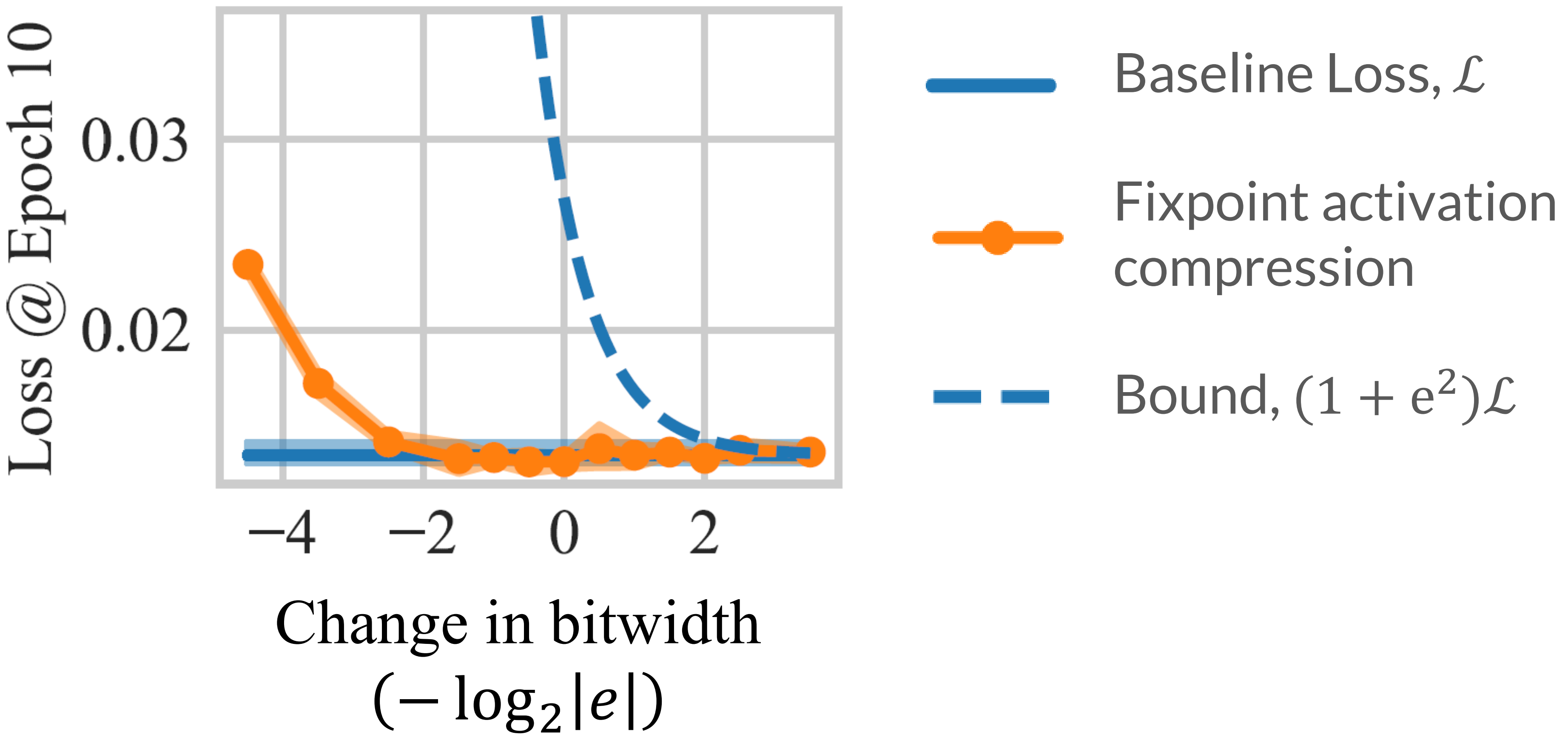
ACCURACY AND COMPRESSION



** Data obtained from original work; may use a different set of models from AC-GC*

(BAA) *A. Chakrabarti, B. Moseley, in NeurIPS 2020*
(ACTNN) *J. Chen, L. Zheng, et. al, in ICML 2021*

THEORETICAL VERSUS EMPIRICAL (MNIST)



Bounds are empirically satisfied

For more information, come to our poster!

AC-GC: Lossy Activation Compression with Guaranteed Convergence

R. David Evans, Tor M. Aamodt

Neural Information Processing Systems (NeurIPS), Dec 7, 2021

Poster Spot A2, 1630 PST – 1800 PST

THANK YOU FOR LISTENING

THE UNIVERSITY OF BRITISH COLUMBIA

Electrical and Computer Engineering

NSERC CRSNG

COHESA