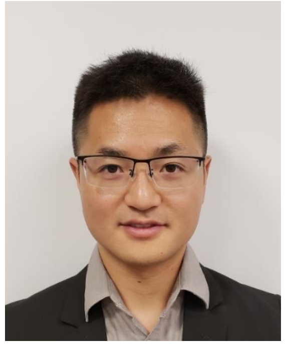
An In-depth Study of Stochastic Backpropagation

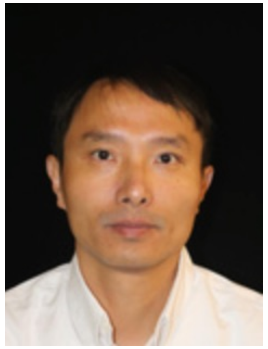
Jun Fang, Mingze Xu, Hao Chen, Bing Shuai, Zhuowen Tu, Joe Tighe

AWS AI Labs

Code: https://github.com/amazon-research/stochastic-backpropagation








Outline

- Motivation
- Method
- Design Strategies
- Generalizability

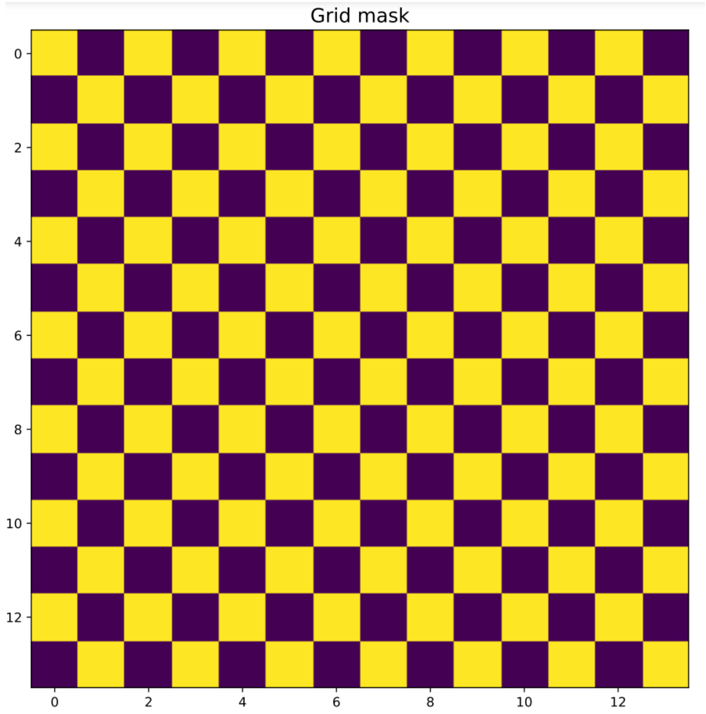
Motivation

- Scaling up models
  - gains higher accuracy
  - but requires more GPU memory usage

- Stochastic Backpropagation (SBP)
  - is a memory efficient training method
  - saves up to 40% of GPU memory for image recognition

Stochastic Backpropagation (SBP)

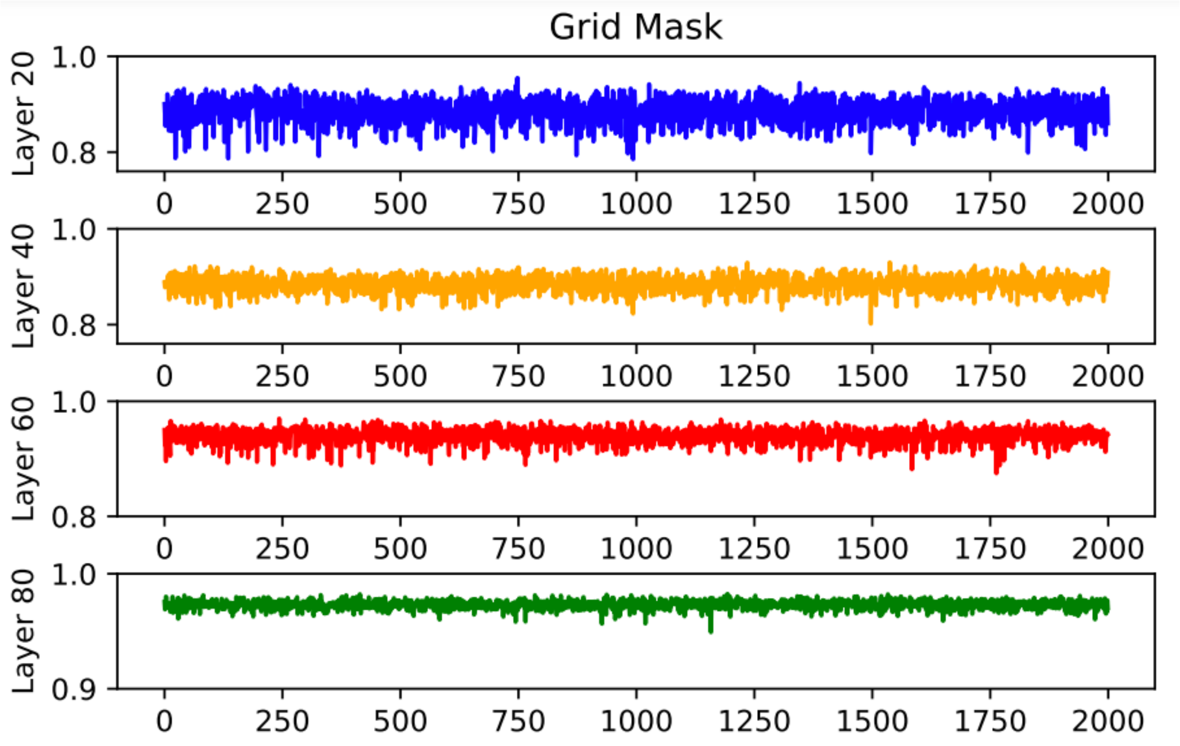
- SBP calculates gradients by **using only a subset of feature maps**



Feature maps at yellow locations are used for SBP gradient calculation

Stochastic Backpropagation (SBP)

- dW (standard SGD) = dW^{keep} (SBP) + dW^{drop}
- We observe that dW^{keep} (SBP) are **highly correlated** (measured by cosine similarity) with dW (SGD)



Implementation

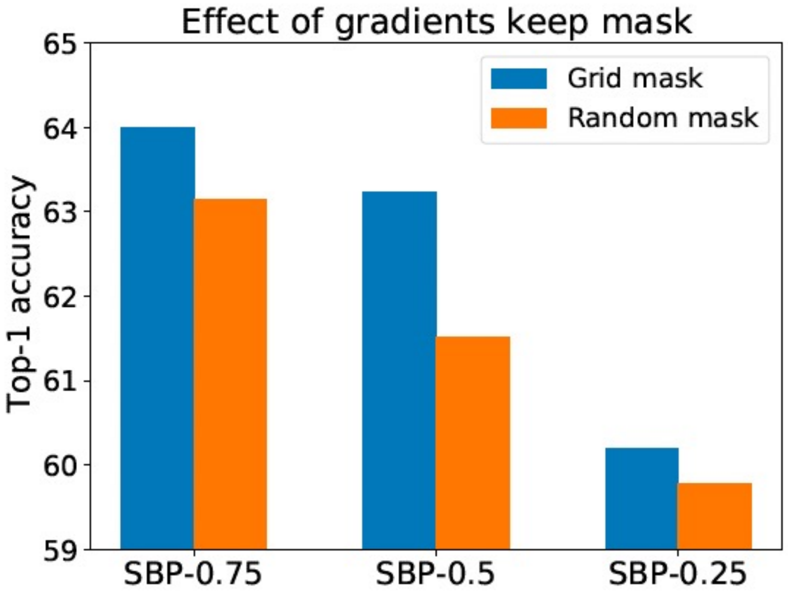
Algorithm 1 Pytorch-like pseudocode of SBP for an arbitrary operation f.

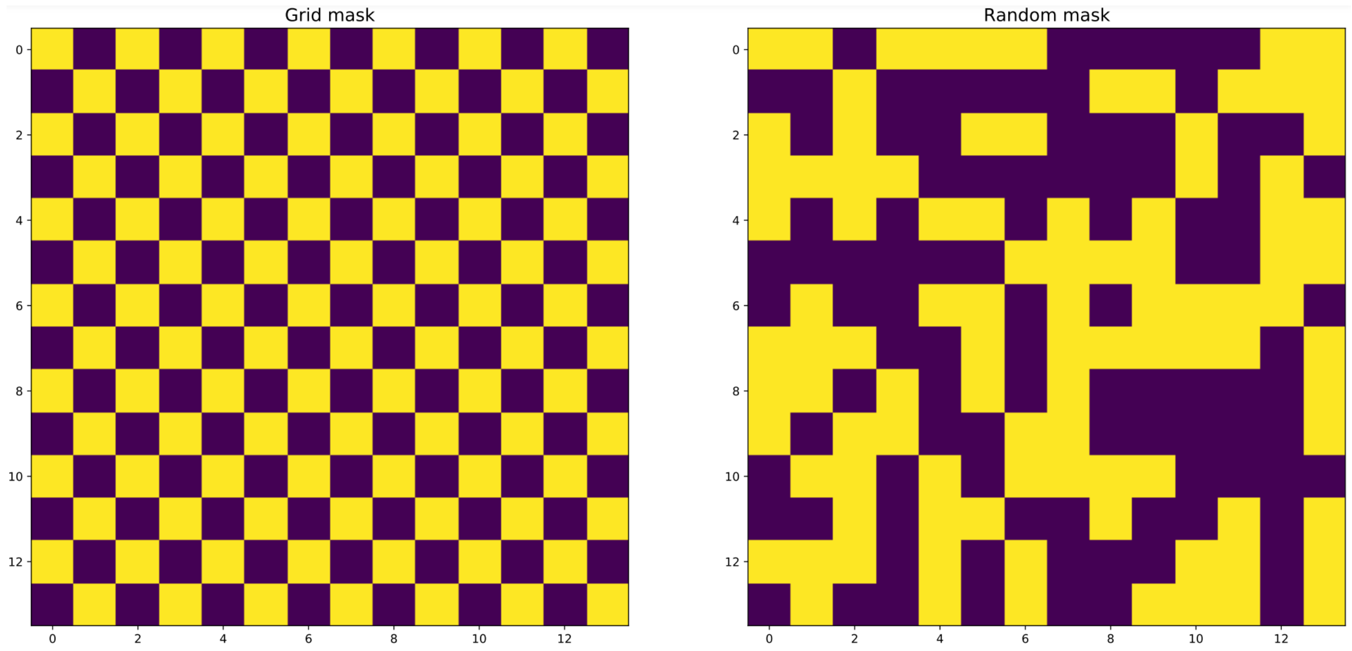
```
# f: an arbitrary operation
# grad_keep_idx: sampled indices where gradients are kept
# grad_drop_idx: sampled indices where gradients are dropped

def sbp_f(f, inputs, grad_keep_idx, grad_drop_idx):
    # initiate outputs
    outputs = torch.zeros(output_shape, device=inputs.device)
    # forward with gradient calculation, gradients will be calculated with torch.autograd
    with torch.enable_grad():
        outputs[grad_keep_idx] = f(inputs[grad_keep_idx])
    # forward without gradient calculation
    with torch.no_grad():
        outputs[grad_drop_idx] = f(inputs[grad_drop_idx])
    return outputs
```

Design Strategies – Gradient Keep Mask

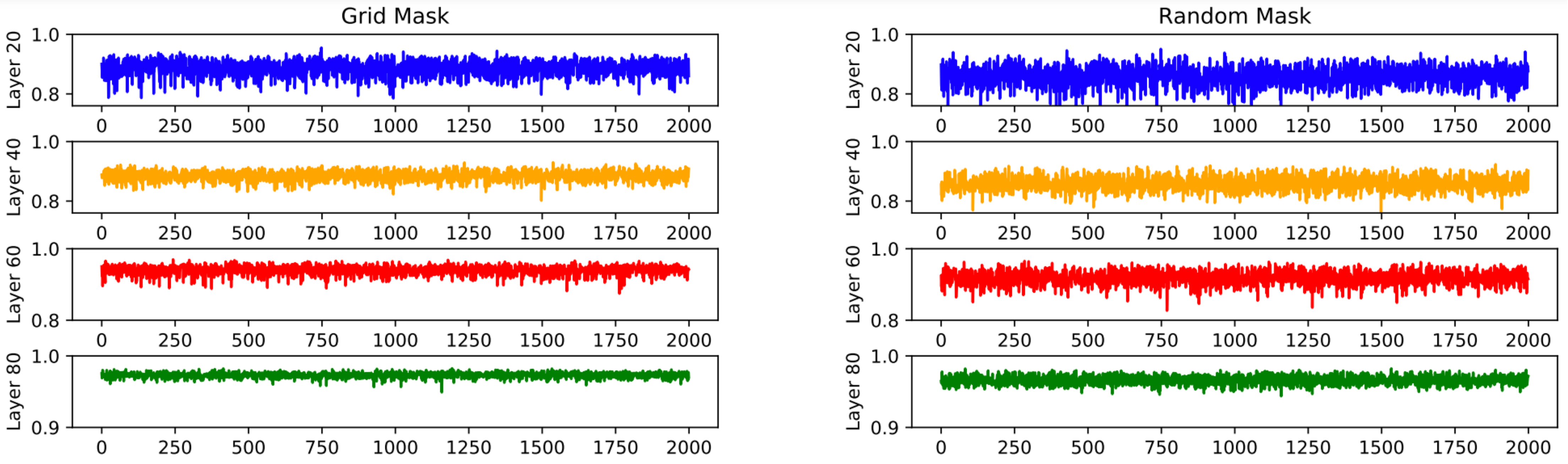
- Grid-wise mask has higher accuracy





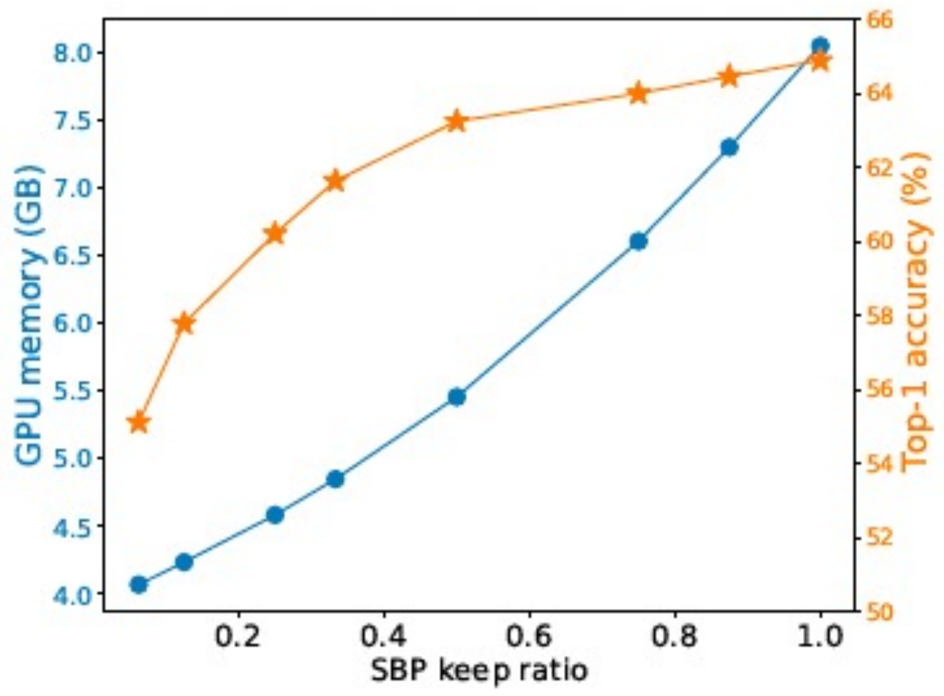
Design Strategies – Gradient Keep Mask

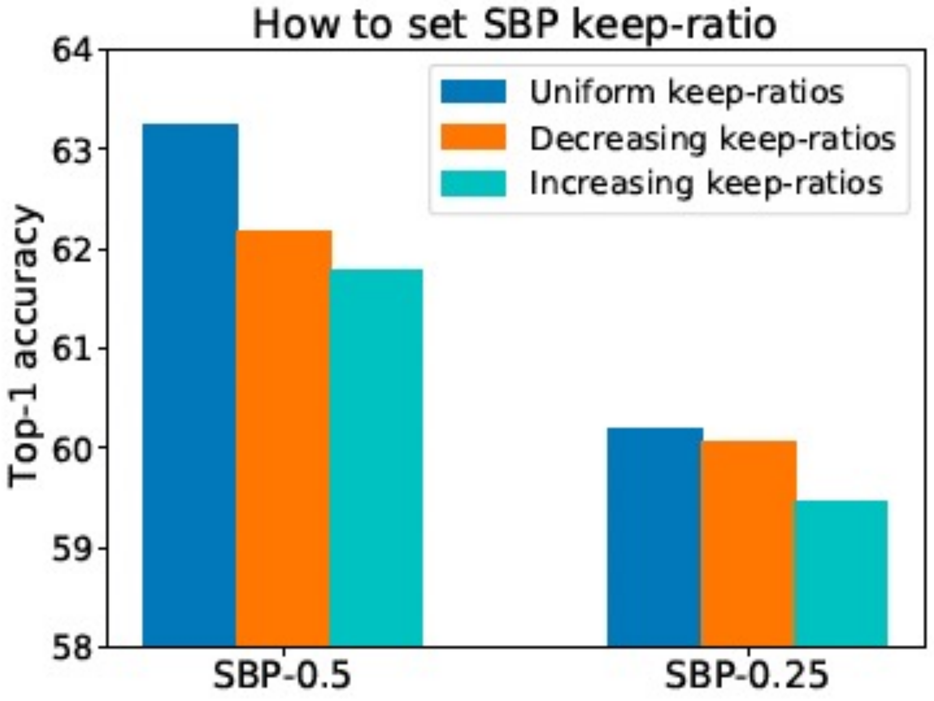
- Grid-wise mask has stronger correlation



Design Strategies – Keep-ratio

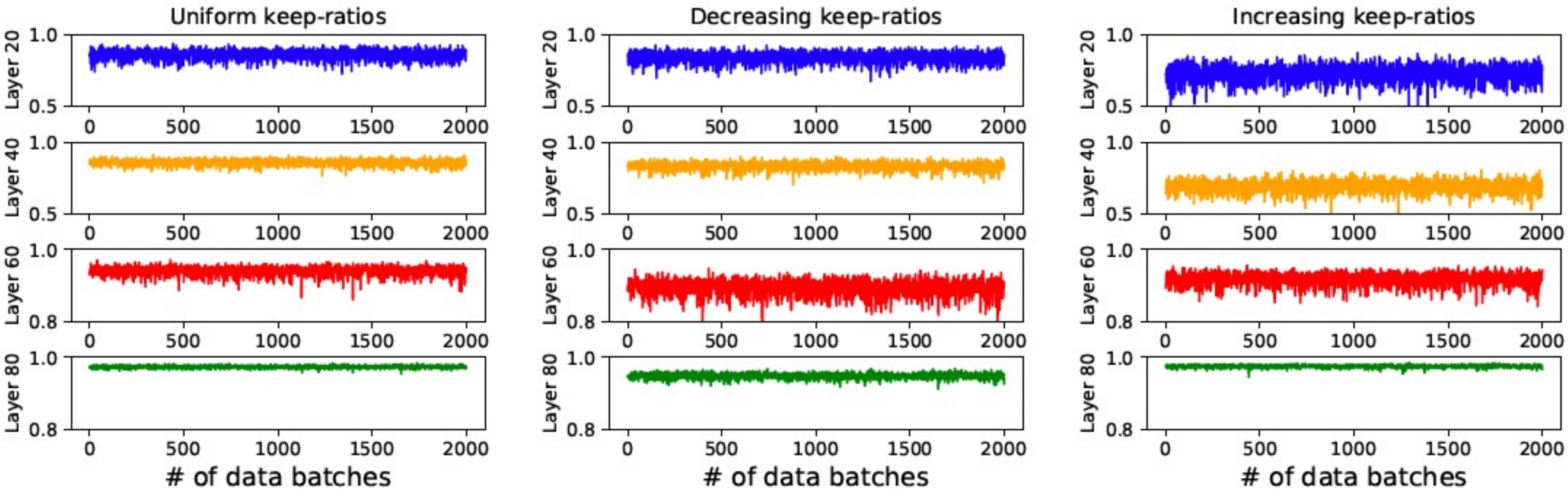
- Sweet spot at keep-ratio = 0.5





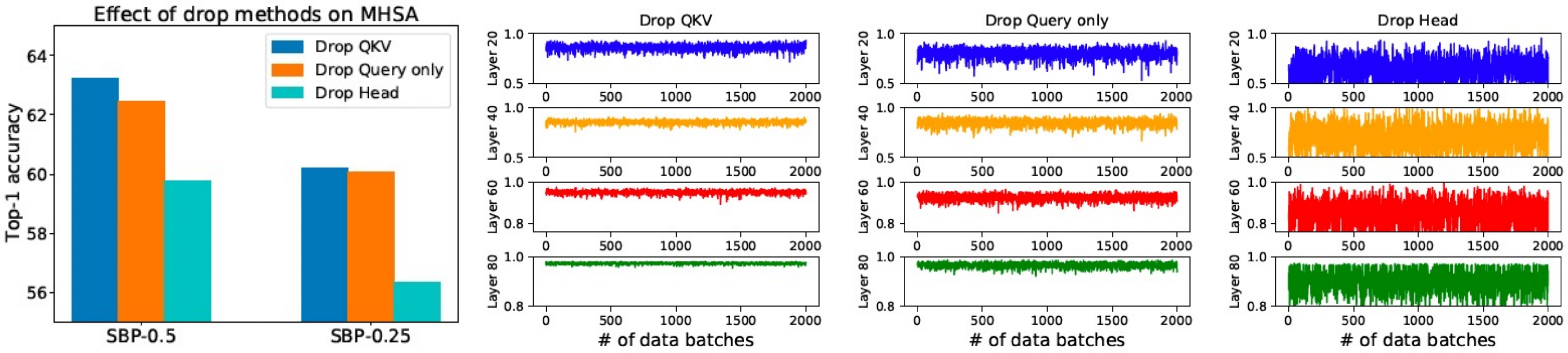
Design Strategies – Keep-ratio

- Uniform keep-ratios method has best accuracy and correlation



Design Strategies – Drop Method on MHSA

- Dropping gradients on all QKV has best accuracy and correlation



Generalizability – ImageNet Classification

- SBP can save up to 40% of GPU memory with 0.6% of accuracy drop

Table 1: Accuracy and memory results of applying SBP for ViT and ConvNeXt on ImageNet.

Network	Keep-ratio	Batch size	Memory (MB / GPU)	Top-1 accuracy (%)
ViT-Tiny	no SBP	256	8248	73.68
ViT-Tiny	0.5	256	5587 (0.68×)	73.09 (-0.59)
ViT-Base	no SBP	64	10083	81.22
ViT-Base	0.5	64	7436 (0.74×)	80.62 (-0.60)
ConvNeXt-Tiny	no SBP	128	12134	82.1
ConvNeXt-Tiny	0.5	128	7059 (0.58×)	81.61 (-0.49)
ConvNeXt-Base	no SBP	64	14130	83.8
ConvNeXt-Base	0.5	64	8758 (0.62×)	83.27 (-0.53)

Generalizability – COCO Object Detection

- SBP can save 30% of GPU memory with 0.7% of accuracy drop

Table 2: COCO object detection and segmentation results using Mask-RCNN with backbone ConvNeXt-T and Cascade Mask-RCNN with backbone ConvNeXt-B.

Backbone	Keep-ratio	Batch size	Memory (GB / GPU)	AP^{box}	AP^{box}_{50}	AP^{box}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
ConvNeXt-T	no SBP	2	8.6	46.2	67.9	50.8	41.7	65.0	44.9
ConvNeXt-T	0.5	2	5.9 (0.69×)	45.5	67.4	50.1	41.1	64.4	44.1
ConvNeXt-B	no SBP	2	17.4	52.7	71.3	57.2	45.6	68.9	49.5
ConvNeXt-B	0.5	2	12.5 (0.72×)	52.5	71.3	57.2	45.4	68.7	49.2

Summary

- SBP calculates gradients by using only a subset of feature maps
- Design Strategies
  - Gradient keep mask
  - Gradient keep ratio
  - Gradient drop method on MHSA
- Generalizability
  - Image classification
  - Object detection
  - Save up to 40% of GPU memory

Thank you!

- Code: https://github.com/amazon-research/stochastic-backpropagation
- Please reach out to junfa@amazon.com if you have any questions!