An In-depth Study of Stochastic Backpropagation

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we provide an in-depth study of Stochastic Backpropagation (SBP) when training deep neural networks for standard image classification and object detection tasks. During backward propagation, SBP calculates the gradients by only using a subset of the feature maps to save memory and the computation cost. We interpret SBP as an efficient way to implement stochastic gradient decent by performing backpropagation dropout, which results in considerable memory saving and training process speed up with a minimal impact to the overall model accuracy. We offer good practice to applying SBP in training deep image classification models, which can be adopted in learning a wide range of deep neural networks. Experiments on image classification and object detection show that SBP can save up to 40% of GPU memory with less than 1% accuracy degradation.

1 Introduction

A common practice to improve accuracy when training deep networks is to increase the input resolution [20], model depth [8], and width [23]. This can significantly increases the GPU memory usage, making training on devices with limited resources difficult. For example, when training an object detector using ConvNeXt-Base [14] as backbone with batch size 2, it requires over 17 GB memory which will not fit on most modern GPUs.

Many prior works have focused on the problem of designing memory efficient training method, especially for large model training on high resolution images or videos. Mixed-precision training [17] uses lower precision to represent the network weights and activations for certain layers of the network. Gradient checkpointing [3] restores intermediate feature maps by recomputing nodes in the computation graph during the backward pass, which has the extra cost effectively performing the forward computation twice. Gradient accumulation splits a mini-batch into smaller chunks to calculate gradients iteratively, which slows down the training process.

A recently developed method, Stochastic Backpropagation (SBP) [4], provides a new angle on memory efficient training. It performs partial execution of gradients during backpropagation on selected frames when training video models [1, 19]. It attains considerable memory saving by only caching a subset of frame feature maps without significant accuracy loss. However, [4] only applies SBP to video based models of two architectures and does not thoroughly explore the design space of the method. The explanation provided in the paper for the effectiveness of SBP is simple but narrow – it attributes the validity of the approximation to the high redundancy of video.

This leaves many questions around SBP unanswered. Here we ask the questions: (1) Is SBP's effectiveness limited to operating on the redundancy of video frames or can it effectively leverage spatial redundancy? (2) If so can, SBP be generalized to image based tasks, with more limited redundancy? (3) Is SBP generalizable to different networks (e.g., MLP, CNN, Transformer) and operators? (4) What are the key design choices when implementing SBP?

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

To answer these questions, we first formalize the SBP process in a general form and analyze the effect SBP has on the calculation of gradients. With this formulation we show how to generalize the SBP idea to image models and present an in-depth study of the design choices available when using SBP. During backward propagation, SBP calculates the gradients from a subset of spatial feature maps to save memory and computation cost. We observe that the calculated gradients by SBP are highly correlated with the standard SGD gradients, indicating SBP is a reasonable approximation. We also provide a new, straightforward implementation of the idea that requires only a few lines of code. We explore several important design strategies which have a noticeable impact on the performance. We validate the generalizability of SBP on two common tasks (image classificaton and detection) with two popular network architectures: ViT [7] and ConvNeXt [14]. We show SBP can effectively train image tasks with less than 1% loss of accuracy in a memory efficient fashion.

2 Understanding Stochastic Back-Propagation (SBP)

In this section, we first formalize the process of Stochastic Back-Propagation (SBP) (sec. 2.1) and show how to compute the gradient for two common operations (sec. 2.2). With these formulations established, we analyze how to perform the chain rule of gradient calculations for SBP (sec. 2.3). Finally, we present a simple implementation of SBP in practice with only a few lines of code (sec. 2.4).

2.1 Formulation of SBP

In this section, we present a general fomulation of SBP [4]. We denote a loss function $L(\Theta)$ to learn a model on a dataset $D = \{s_i\}_{i=1}^{N}$, where Θ is the collection (vector, matrix, or tensor) of all parameters of the model, N is the total number of training samples, $s_i \in \mathbb{R}^{T\times H\times W\times C}$ is an input data sample, where H and W are spatial resolution height and width, C is the input channel size, $T = 1$ represents an image and $T > 1$ represents a video. In each training iteration, the model with SBP processes a mini-batch of data samples $S \in \mathbb{R}^{B\times T\times H\times W\times C} \subseteq D$ to compute the loss in the forward pass, where B is the batch size. During the backward pass, differing from the traditional back-propagation that relies on the full feature maps $X_i \in \mathbb{R}^{B\times T_i\times H_i\times W_i\times C_i}$($i$ is the index of the layers) an incomplete execution for back-propagation (i.e., SBP) only utilizes a subset of the feature maps $X_i^{sub} \in \mathbb{R}^{T_i^{sub}\times H_i^{sub}\times W_i^{sub}\times C_i}$ to approximate the gradient calculation but keep all the full forward paths. That is, it updates the model weights as

$$\Theta_{t+1} = \Theta_t - \eta\frac{1}{B}\sum_{X_i^{sub}\in F}\nabla L(X_i^{sub}, \Theta_t), \tag{1}$$

where η is the learning rate. Compared to the traditional back-propagation that uses the full feature maps, SBP saves memory by only caching a subset of feature maps. Note that the method in [4] is a special case of (1) as it only performs backward calculation on a subset of video frames $X_i^{T_i^{sub}} \in \mathbb{R}^{T_i^{sub}\times H_i\times W_i\times C_i}$.

2.2 Backward Gradient Calculation

Next, we derive the equations to compute the gradients during the backwards phase of SBP. We denote that at layer i of the model, an operator f_i with learnable weights W_{f_i} to processes the feature maps $X_i \in \mathbb{R}^{B\times T_i\times H_i\times W_i\times C_i^{in}}$, where C_i^{in} is the input channel size.

2.2.1 Calculation of Linear Layers

We start with exploring f_i as a point-wise convolutional layer (i.e., PW-Conv). PW-Conv is widely used in the state-of-the-art computer vision models such as ViT [7], ConvNeXt [14], and Video Swin [13]. It is equivalent to the linear layers in Transformer MLP blocks [14], and is also known as channel-wise fully connection layer. It can be either implemented through the linear layer, or the 1×1 convolution. It is mainly used to enrich the channel information with weights $W_{f_i} \in \mathbb{R}^{C_i^{in}\times C_i^{out}}$, and the analysis on PW-Conv can be easily extended to other layers. The forward pass of this layer is (the bias term is omitted):

$$X_{i+1} = f_i(X_i) = X_i W_{f_i}, \tag{2}$$

where $X_{i+1} \in \mathbb{R}^{B\times T_i\times H_i\times W_i\times C_i^{out}}$. During the traditional backward pass with full back-propagation, the gradients of both activations and weights are calculated by:

$$\begin{aligned} dX_{i+1} &= \partial Loss/\partial X_{i+1}, \\ dW_{f_i} &= \partial Loss/\partial W_{f_i} = X_i^T dX_{i+1}, \\ dX_i &= \partial Loss/\partial X_i = dX_{i+1} W_{f_i}^T. \end{aligned} \tag{3}$$

It requires to store the weights $W_{f_i} \in \mathbb{R}^{C_i^{in} \times C_i^{out}}$ and the full feature map $X_i \in \mathbb{R}^{B \times T_i \times H_i \times W_i \times C_i^{in}}$ to compute the gradients. In many cases, $B \times T_i \times H_i \times W_i \gg C_i^{out}$, i.e., caching of feature maps dominants the memory usage. In order to effectively save GPU memory during training, SBP only caches a subset of the feature map for back-propagation.

In SBP, we split the full feature map X_i into two subsets X_i^{keep} and X_i^{drop}, to denote the corresponding indices of the sampled feature maps to be kept and dropped for the backward gradients calculation. An example of X_i^{keep} can be $X_i[:, \text{even index on } T, :, :, :]$ for video, or $X_i[:, :, \text{even index on } H, \text{odd index on } W, :]$ for image (e.g., Fig. 3 left), and X_i^{drop} is the complementary part of the feature map. We have $X_i^{keep} \cap X_i^{drop} = \emptyset$ and $X_i^{keep} \cup X_i^{drop} = X_i$. We denote keep-ratio r as the number of gradients kept indices $j \in \mathbb{Z}_i^{keep}$ over the number of all indices.

One advantage of PW-Conv operator to simplify the SBP problem is, the nodes in spatial or temporal dimensions are calculated independently through the forward and backward passes. Hence, with X_i^{keep} and X_i^{drop} we can update the forward pass separately on the kept and dropped indices:

$$\begin{aligned} &X_{i+1} = [X_{i+1}^{keep}, X_{i+1}^{drop}] = [X_i^{keep}, X_i^{drop}] W_{f_i}, \\ &\text{i.e.,} \quad X_{i+1}^{keep} = X_i^{keep} W_{f_i}, \quad X_{i+1}^{drop} = X_i^{drop} W_{f_i} \end{aligned} \tag{4}$$

as well as the backward:

$$\begin{aligned} &dX_{i+1} = [dX_{i+1}^{keep}, dX_{i+1}^{drop}], \\ &dX_i^{keep} = dX_{i+1}^{keep} W_{f_i}^T, \quad dX_i^{drop} = dX_{i+1}^{drop} W_{f_i}^T, \\ &dW_{f_i} = X_i^T dX_{i+1} = [X_i^{keep}, X_i^{drop}]^T [dX_{i+1}^{keep}, dX_{i+1}^{drop}] \\ &= X_i^{keep^T} dX_{i+1}^{keep} + X_i^{drop^T} dX_{i+1}^{drop}. \end{aligned} \tag{5}$$

SBP drops all the gradients calculation with superscripts ^{drop} and keeps all the gradients calculation with superscripts ^{keep}, thus it saves memory and computation on the dependencies of dropped indices. In other words, the memory usage and the backward computation cost is proportional to the keep-ratio of the original case. Mathematically, it sets $dX_{i+1}^{drop} = 0$ (thus $dX_i^{drop} = 0$) and uses $dW_{f_i}^{keep} = X_i^{keep^T} dX_{i+1}^{keep}$ as a stochastic approximation of the original weight gradients dW_{f_i} to update the model weights.

2.2.2 Calculation of Convolutional Layers

We then derive the calculation of a general convolutional layer as f_i. SBP of convolutional layer is more involved as unlike the PW-Conv, the nodes in spatial or temporal dimensions are no longer independent in forward and backward calculation. Assume that we apply a convolutional kernel $k \times k$ with stride s. The regular back-propagation of the convolution layer is calculated as follows:

$$\begin{aligned} &dX_i = zeropad(dX_{i+1}) * W'_{f_i} \\ &dW_{f_i} = X_i * dX_{i+1} \end{aligned} \tag{6}$$

where $*$ is the convolution operation, $zeropad$ is to pad the edge of the matrix with zeros, and W'_{f_i} represents the $180°$ rotation of W_{f_i}.

Due to the convolution operator in the gradient calculation, given a location j on the input feature map as $x_i^j \in X_i$, all the elements from dX_{i+1} inside the region of the kernel size k would contribute to the gradient calculation of element x_i^j. Therefore, in SBP of convolutional layers, when we set $dX_{i+1}^{drop} = 0$ for certain areas, in general, two things would happen: (a) even if $j \in \mathbb{Z}^{drop}$ and $dx_{i+1}^j = 0$, dx_i^j would not be 0 due to the contribution of the neighbors of j on dX_{i+1}; (b) dx_i^j would be an approximate gradient unless all dx_{i+1}^p with p in the neighbor areas of j in terms of the kernel size k are kept and there is no chain rule effects on dx_{i+1}^p. However, there is a special scenario, i.e., when $s \geq k$, in which the neighbor effects disappear due to the fact that dX_{i+1} needs to be zero interweaved in gradient calculation. In such a scenario, nodes in spatial or temporal dimensions are getting more independent and the scenario is getting similar to that of PW-Conv.

2.2.3 Discussion

From the derivation above, we see that the SBP approximation behaves differently depending on the operator. For layers such as linear layer or PW-Conv, the gradients (i.e., dx_i) are either 0 or exact[1], while for convolutional layers with kernel size larger than 1, most of the gradients on the activations are neither 0 nor exact. Due to space limitation, we only derived SBP on two most popular operators. For derivation of other operators please refer to the supplemental material.

2.3 Chain Rule Effect

We have derived the calculation of SBP on a single layer in previous section. In this section, we derive the calculation of SBP on a stack of multiple layers. For simplicity, we take two layers as an example, which can be easily generalized to more layers.

2.3.1 Two PW-Conv Layers

We first consider a two PW-Conv layer scenario. We assume a layer before f_i, i.e., f_{i-1}, is another PW-Conv layer with weights $W_{f_{i-1}}$. The forward pass of this layer is (the bias term is also omitted):

$$X_i = f_{i-1}(X_{i-1}) = X_{i-1}W_{f_{i-1}}. \tag{7}$$

Then we can calculate the gradients of weights and activations in layer f_{i-1} similar to Eq. (3):

$$\begin{aligned} dX_{i-1} &= dX_i W_{f_{i-1}}^T = dX_{i+1}W_{f_i}^T W_{f_{i-1}}^T, \\ dW_{f_{i-1}} &= X_{i-1}^T dX_i = X_{i-1}^T dX_{i+1} W_{f_i}^T. \end{aligned} \tag{8}$$

We first consider the case that there is no additional gradient dropout at layer f_{i-1}. Since dX_{i+1} has gradients at kept indices (i.e., dX_{i+1}^{keep}) and 0 at dropped indices (i.e., dX_{i+1}^{drop}), Eq. (8) can be updated to:

$$\begin{aligned} &dX_{i-1} = [dX_{i-1}^{keep}, dX_{i-1}^{drop}], \\ &dX_{i-1}^{keep} = dX_{i+1}^{keep} W_{f_i}^T W_{f_{i-1}}^T, \quad dX_{i-1}^{drop} = dX_{i+1}^{drop} W_{f_i}^T W_{f_{i-1}}^T = 0, \\ &dW_{f_{i-1}} = {X_{i-1}^{keep}}^T dX_{i+1}^{keep} W_{f_i}^T. \end{aligned} \tag{9}$$

From the equation we can see that if there is no additional gradient dropout applied on layer f_{i-1}, through the chain rule, the keep and drop set would transfer from X_{i+1} to X_{i-1} identically.

If we apply gradient dropout on layer f_{i-1} as well, i.e., on layer f_{i-1} we split dX_i into dX_i^{keep} and dX_i^{drop}. Assume the keep / drop indices set at layer $i+1$ is $\mathbb{Z}_{i+1}^{keep}$ and $\mathbb{Z}_{i+1}^{drop}$, and at layer i is $\mathbb{Z}_i^{keep}$ and $\mathbb{Z}_i^{drop}$, then through Eq. (8) and Eq. (9), the keep / drop indices set on layer $i-1$ would be:

$$\mathbb{Z}_{i-1}^{keep} = \mathbb{Z}_{i+1}^{keep} \cap \mathbb{Z}_i^{keep}, \quad \mathbb{Z}_{i-1}^{drop} = \mathbb{Z}_{i+1}^{drop} \cup \mathbb{Z}_i^{drop}. \tag{10}$$

Eq. (9) becomes:

$$\begin{aligned} &dX_{i-1} = [dX_{i-1}^{keep'}, dX_{i-1}^{drop'}], \\ &dX_{i-1}^{keep'} = dX_{i+1}^{keep'} W_{f_i}^T W_{f_{i-1}}^T, \quad dX_{i-1}^{drop'} = 0, \\ &dW_{f_{i-1}} = {X_{i-1}^{keep'}}^T dX_{i+1}^{keep'} W_{f_i}^T. \end{aligned} \tag{11}$$

where $keep' = \mathbb{Z}_{i-1}^{keep}$ and $drop' = \mathbb{Z}_{i-1}^{drop}$.

2.3.2 PW-Conv + Conv Layers

Now let's look at the chain-rule effect of the case when f_{i-1} is a convolutional layer with kernel size k > 1. Assume that we apply a convolutional kernel $k \times k$ with stride s. Using the chain rule we can update the Eq. (6) as follows:

$$\begin{aligned} dX_{i-1} &= zeropad(dX_i) * W'_{f_{i-1}} = zeropad(dX_{i+1}W_{f_i}^T) * W'_{f_{i-1}}, \\ dW_{f_{i-1}} &= X_{i-1} * dX_i = X_{i-1} * dX_{i+1} W_{f_i}^T. \end{aligned} \tag{12}$$

[1] by "exact" we mean the exact calculation of the gradient for the given mini-batch, though it is still an estimation due to SGD.

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
```

We can see from the equation that each dx_{i-1} receives the backward gradients from the outputs of its neighbors, which might be dropped out and thus the gradient at dx_{i-1} are no longer the exact.

From the Eq. (12). The dropout at layer i makes the matrix dX_i sparse (only indices at X_{i+1}^{keep} are non-zero). If we conduct dropout at the current layer f_{i-1} (i.e., we have dX_i^{keep} and dX_i^{drop}), the matrix dX_i will become even more sparse unless the dropout is happening on the same location. Stacking layers further and further in this way, the chain rule effect may lead towards a very sparse output derivative matrix (e.g., dX_i), in which the nodes are getting more spatial or temporal independent in gradient calculation like what is in the PW-Conv operator.

However, stacking convolutional layers (due to the space limitation, we skip the math derivation) would not have such trend, as the gradient dropout in convolutional layer only creates approximate gradient but not many zeros, due to the neighbor effects in calculation.

2.3.3 Discussion

Overall, in terms of the chain rule effect, we can see from Eq. (11) that stacking PW-Conv layers, especially in an interweaving way may lead to a sparse keep set. However, if we always drop gradient at the fix position then this won't happen. Stacking PW-Conv + Conv layers (in Eq. (12)) also tends to make the keep set sparse, while stacking Conv layers doesn't have such effect, though it stacks the non-exact gradient at almost every position and tends to make the gradients gradually move away from the original ones.

2.4 Efficient Implementation

With a deeper understanding of how SBP works, we provide a simple and efficient implementation of the SBP technique in Alg. 1. We point out that this is more efficient than the prior [4] work as there is (1) no need to do the re-forward in the backward so it is computationally more efficient, and (2) no need to manually cache the intermediate feature maps and fill zero gradients in the backward, which is much simpler. We emphasize that it is very simple and easy to implement, it only needs 3 lines of code to do the forward, and the backward (no need to implement manually) can be seamlessly taken care of by the ease of "autograd engine" in the deep learning frameworks. It supports point-wise convolutional layers, the dot-product attention and multilayer perception (MLP) in popular networks including ViT [7] and its variants [7, 12, 13], ConvNeXt [14] and hybrid architectures [5].

3 Design Strategies

In this section, we investigate the effectiveness of different design strategies of SBP on image tasks. We use ViT-Tiny [7] model to evaluate the top-1 accuracy on ImageNet [18] validation dataset. The ViT-Tiny is a smaller version of the ViT [7] variants, it uses patch size 16 with 12 transformer layers, the embedding dimension is 192 and number of heads is 3, i.e., each head has dimension 64. The experiments are trained for 100 epochs on the ImageNet training dataset. The conclusions remain for larger models and higher training epochs.

How many layers should be applied with SBP? To answer this question, we progressively apply SBP to transformer layers from the last layers to early layers and present their results in Fig. 1 Left. As shown, we observe a sweet spot when applying SBP to 8 out of 12 layers, which saves 33% of GPU memory with an accuracy drop within 1.5%. Although applying SBP on all the transformer

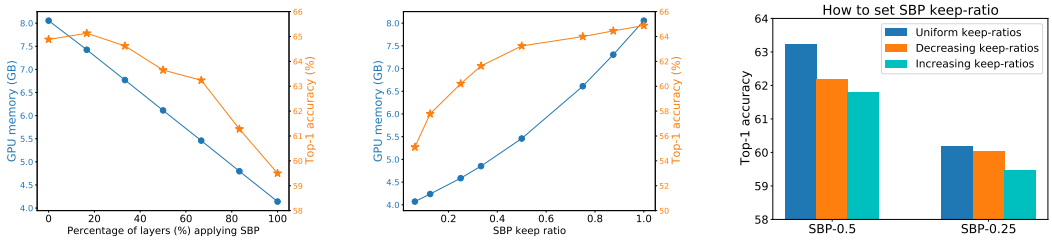

Figure 1: Illustration of the effect on applying SBP on different choices of the network layers and keep-ratios. Left: accuracy curve of applying SBP on different percentage of layers, the keep ratio of SBP is 0.5. Middle: trade-off of keep-ratio versus accuracy, every SBP layer has the same keep-ratio. Right: comparison of using uniform keep-ratios, progressively increasing and decreasing keep-ratios for the SBP layers.

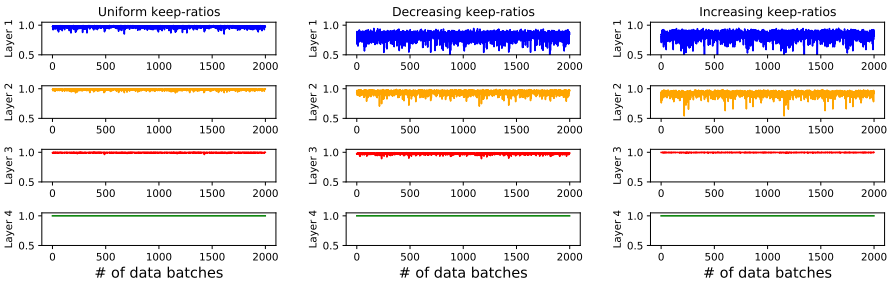

Figure 2: Cosine similarity of weights gradients for applying SBP on linear Layer1, 2, and 3 with a uniform (0.5-0.5-0.5), decreasing (0.75-0.5-0.25), or increasing (0.25-0.5-0.75) keep-ratios.

layers can save the most GPU memory, it leads to a large accuracy drop. Therefore, unless stated otherwise, we apply SBP on 2/3 of all layers of ViT models for the rest of this work.

How much gradients should be kept in SBP? In this section, we discuss the trade-off between the gradients keep-ratio and accuracy in SBP. As a reminder, keep-ratio r is the number of gradients kept indices over the number of all indices, and it represents the percentage of spatial feature map values to calculate the gradients. In general, the higher the keep-ratio, the more gradient information is preserved, and the higher the accuracy, but the less is memory saved. A natural question is what is a good keep-ratio to balance the accuracy drop and memory saving.

In Fig. 1 Middle, we show the model's accuracy with respect to the keep-ratio of gradients. In this case, we apply the same keep-ratio to all the SBP layers at the same gradients kept indices and we call it uniform keep-ratios. We observe that setting keep-ratio to be 0.5 is a golden number, and having keep-ratio to be 0.25 is too aggressive, which results in a dramatic accuracy drop (3.04%) of the model. We also explore progressive keep-ratios, which we linearly increase / decrease the keep-ratios from early to late layers while keeping their average keep-ratio to be 0.5. In detail, the keep-ratios are $[0.25, 0.32, 0.39, 0.46, 0.53, 0.60, 0.68, 0.75]$ or its inverse on 8 layers. As shown in Fig. 1 Right, both models perform significantly worse.

We simulate the gradient calculation on a 4 layer network (Conv1 - Conv2 - Linear3 - Linear4 - Loss). We apply different keep-ratio methods on the first 3 layers, and calculate the correlation (measured by cosine similarity) of the gradients weights of applying SBP to the original exact gradients without applying SBP. From Fig. 2, we observe that the weights gradients have a stronger correlation by using uniform keep-ratios compared to increasing or decreasing keep-ratios. We believe uniform keep-ratios can keep the consistency of gradient dropping locations between different layers and can preserve gradient information at these spatial locations. However, non-uniform (progressive increasing or decreasing) keep-ratios will have different dropping locations between different layers. As an overall result, it may loss more gradient information on more spatial locations and produce a less accurate estimation of the original gradients and may have a lower accuracy performance.

How to sample the gradient keep mask? SBP calculates the gradients with only using a subset of feature maps, a key question is how to sample the subset of feature maps to mask the kept gradients. If we fix the spatial locations of the gradient kept indexes, then the backward only propagates the valid gradients (non-zeros) on these fixed locations, which force the model only learns from fixed partial spatial information and therefore severely hurts the model performance. A better strategy is to

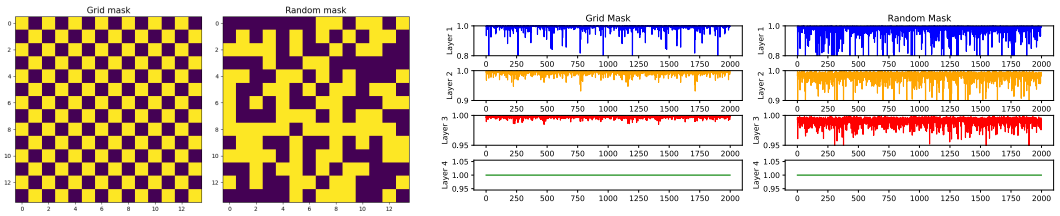

Figure 3: Left two: a particular example of gradient keep mask for grid-wise sampling and random sampling. Right two: cosine similarity of weights gradients for applying SBP-0.5 on grid-wise sampling mask and random sampling mask. SBP is applied on Layer3.

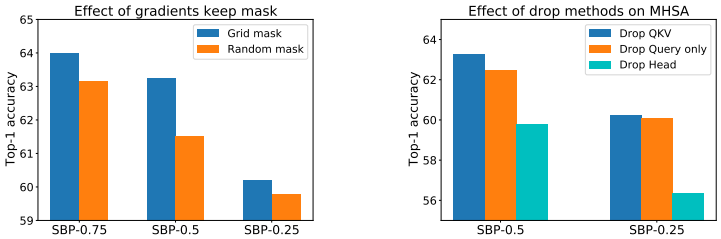

Figure 4: Left: effect of different sampling method for gradient keep mask. Right: effect of different SBP drop methods on Multi-Head Self-Attention.

randomly alternate the gradient kept indexes in each training step, such that it can statistically visit every spatial location.

We examine two mask sampling strategies: grid-wise mask sampling and random mask sampling, see Fig. 3 left as a particular example of keep-ratio 0.5. We report the accuracy in Fig. 4 left. Overall, grid-wise mask sampling achieves $0.5\% \sim 1.7\%$ higher accuracy over random mask sampling. In Fig. 3, we also observe that grid-wise sampling has a stronger correlation for weights gradients. We believe this is because grid-wise sampling is a more structured sampling method and it provides a more accurate estimation of gradients calculation. Unless stated otherwise, we use grid-wise sampling in this paper.

How to apply SBP on MHSA? In general, transformer block consists of two sub-blocks: Multi-Head Self-Attention (MHSA) and Multi-layer perception (MLP). Theoretically, the memory usage of applying SBP on MLP (two linear layers) is proportional to the keep-ratio r. However, the memory saving of applying SBP on MHSA depends on the sampling method on query, key, value (QKV) and the ratio of head dimension d over the number of tokens n. From [4], the activation maps needed for gradients calculation in MHSA includes the inputs and outputs ($2hdn$), QKV vectors (each with hdn), the attention weight maps ($2hnn$ before and after the softmax), thus requires the memory of $3hdn + 2hnn$ in total, where h is the number of heads. Dropping gradients on query only (Drop Query) or on all QKV (Drop QKV) enjoys the following memory usage ratio:

$$\begin{aligned} &\text{Query only:} \quad \frac{2hdn+hdnr+2hnnr}{3hdn+2hnn} = \frac{r(\frac{d}{n}+2)+2\frac{d}{n}}{(\frac{d}{n}+2)+2\frac{d}{n}} \\ &\text{QKV:} \quad \frac{3hdnr+2hnnr^2}{3hdn+2hnn} = r\frac{3+2r\frac{d}{n}}{3+2\frac{d}{n}} \end{aligned} \tag{13}$$

In video transformers, the number of tokens n are much larger than the head dimension d (i.e., in video Swin Transformer, $n = 392$ and $d = 32$, $\frac{d}{n} = 0.082$) and thus the attention weight maps $2hnn$ comprise most of the memory. Therefore, dropping gradients on query only in video model is good enough to save memory such that its memory is linearly proportional to the keep ratio r [4].

However, in image task models such as ViT with a typical input size 224 and patch size 16, $n = 14 \times 14 = 196$ and $d = 64$, $\frac{d}{n} = 0.327$. Then with keep-ratio $r = 0.5$ (or $r = 0.25$), dropping gradients on query alone uses $0.61\times$ (or $0.42\times$) memory while dropping all QKV reduces its memory usage ratio to $0.46\times$ (or $0.22\times$). From Fig. 4, dropping on all QKV also gains better accuracy. We also report the performance of dropping gradients along the attention heads but did not observe accuracy improvement. Hence, we apply SBP on all QKV for image-based transformers.

We plot the cosine similarity of the weights gradients of applying SBP and the original exact gradients in Fig. 5. One interesting observation is that Drop QKV has a slightly better correlation to Drop Query

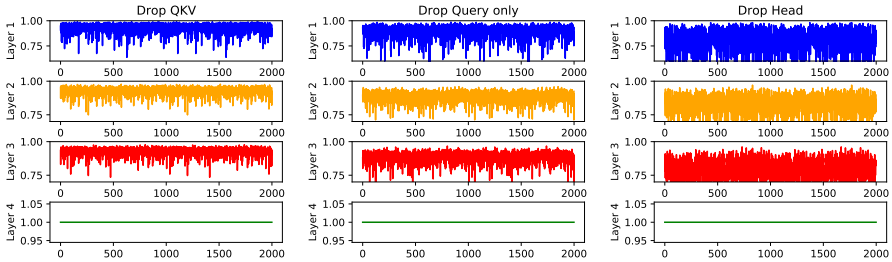

Figure 5: Cosine similarity of weights gradients for different method on applying SBP-0.5 for MHSA. The layer 3 is the MHSA layer applied with SBP with a keep-ratio of 0.5. Left: Drop QKV, middle: Drop Query only, right: Drop Head.

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	72.68
ViT-Tiny	0.5	256	5590 (0.68×)	72.64 (-0.04)
ViT-Base	no SBP	64	10083	79.06
ViT-Base	0.5	64	7430 (0.74×)	78.63 (-0.43)
ConvNeXt-Tiny	no SBP	128	12134	82.1
ConvNeXt-Tiny	0.5	128	7115 (0.59×)	81.60 (-0.50)
ConvNeXt-Base	no SBP	64	14130	83.8
ConvNeXt-Base	0.5	64	8758 (0.62×)	83.32 (-0.48)

and a much higher correlation to Drop Head. We believe an accurate estimation of the gradients might be an important factor to the model performance trained with SBP, which provides an insight for further investigation.

4 Generalizability of SBP

We evaluate the generalizability of our proposed SBP on two computer vision benchmarks: image classification on ImageNet [18] and object detection on COCO [11]. In order to do a fair comparison of the same network, we keep the training hyper-parameters of optimizer, augmentation, regularization, batch size and learning rate the same for a given model, and only adopt different stochastic depth augmentation [10] for different model sizes. The Mixed Precision Training [17] method is enabled for faster training. All experiments are conducted on 8× Tesla 16GB V100.

4.1 Classification on ImageNet

We report the memory and accuracy trade-off in Tab. 1 for two state-of-the-art networks: transformer-based ViT [7] and convolutional-based ConvNeXt [14]. We train 300 epochs for both these two networks. We use DeiT's [21] training recipe for ViT and use the official one of ConvNeXt.

For ViT models, we apply SBP on the last 8 transformer blocks including MHSA and MLP layers with the keep-ratio of 0.5, reducing GPU memory usage by approximately 30%. Note that the memory saving ratio of the network is not proportional the the drop-ratio ($1-$ keep-ratio) because we only apply SBP on $\frac{2}{3}$ of all the layers. With keep ratio of 0.5, the SBP technique can still effectively learn with an accuracy drop of only 0.04% and 0.43% for ViT-Tiny and ViT-Base, respectively.

For ConvNeXt models, it builds the basic block with a depth-wise convolutional (DW-Conv) layer with kernel size 7×7 followed by two point-wise convolutional (PW-Conv) layers[2] and the DW-Conv layer only consumes a small portion of GPU memory (less than 10%). It has multi-scale stages with early stages having higher redundancy and occupy most of the GPU memory. Therefore, we apply SBP on the PW-Conv and down-sampling layers in the first three stages, and do not apply it on the last stage. With keep ratio of 0.5, ConvNeXt-Tiny and ConvNeXt-Base can efficiently learn the model with a GPU memory saving of above 40% and an accuracy drop within 0.5%.

[2]The two PW-Conv layers is equivalent to an MLP block in transformers.

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	67.3	49.9	41.0	64.4	43.9
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	71.4	57.2	45.4	68.8	49.2

4.2 Object Detection and Segmentation on COCO

We fine-tune Mask R-CNN [9] and Cascade Mask R-CNN [2] on the COCO dataset [11] with ConvNeXt-Tiny and ConvNeXt-Base backbones pretrained on ImageNet-1K, respectively. Following the same hyper-parameter settings of [14], we apply SBP on the ConvNeXt backbones to save training memory. Results in Tab. 2 show that SBP with keep-ratio of 0.5 can still learn the detection task reasonably well with 0.2% and 0.7% loss on the box and mask average precision (AP) for ConvNeXt-Base and ConvNeXt-Tiny backbones, respectively. However, it only consumes about $0.7\times$ of the GPU memory.

Note that training detection models on high resolution images (up to 800×1333) is memory intensive even for a very small batch size, e.g., with backbone of ConvNeXt-Base and batch size of 2, training requires 17.4GB of GPU memory which may lead to out of memory error for common devices such as Tesla 16GB V100. While with the help of SBP, the model can be still effectively trained on GPUs with 12.5GB memory, which is much more feasible for many research groups. Here we apply SBP only in the backbone, we believe it has a potential to apply SBP on more other layers in the detection model to achieve more promising efficiency results.

5 Related Work

Since deep neural networks are highly extended in both depth [8] and breadth [23] in recent years, designing memory efficient training methods has attracted a fair amount of attention. By saving and recomputing nodes in the computation graph during backpropagation, Gradient Checkpoint [3] can save a large amount of memory (up to about 50%). On the other hand, gradient accumulation splits the large sample batch into several mini batches and runs them iteratively without updating the model parameters. This saves the GPU memory proportionally without affecting the accuracy because accumulating the gradients of these sub-iteration and then updating the model parameters is identical to directly optimizing the model global batch size. However, these two methods are slow down the training process. Multigrid [22] proposes to use different mini-batch shapes to speedup the model training. Sparse Network [6] particularly targets on the recognition tasks but can only save the memory theoretically. Sideways [15] and its follow-up [16] reduce the memory cost by overwriting activations whenever new ones become available but can only be applied to causal models. Stochastic Backpropagation (SBP) [4] reduces a large portion of GPU memory by only backpropagating gradients from incomplete execution.

6 Discussion

In this work, we present a comprehensive study of the Stochastic Back-Propagation (SBP) mechanism with a generalized implementation. We analyze the effect of different design strategies to optimize the trade-off between accuracy and memory. We show that our approach can reduce 40% of the GPU memory when training image classification models under various deep learning backbones.

Limitations: There are several limitations of our current work. In general, SBP is a memory efficient training method, but as byproduct, it still causes slightly loss of accuracy. In our current results, SBP brings only a small amount of training speedup($\sim 1.1\times$), how to further speed up the training process is part of our future work.

Societal impact: Our work studies the stochastic backpropagation mechanism, which can be used for training general deep learning models. Misuse can potentially cause societal harm. However, we believe our work is a general approach and is consist with the common practice in machine learning.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *ICML*, 2021.

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018.

[3] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv:1604.06174*, 2016.

[4] Feng Cheng, Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Li, and Wei Xia. Stochastic backpropagation: A memory efficient strategy for training video models. *CVPR*, 2022.

[5] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 2021.

[6] Tim Dettmers and Luke Zettlemoyer. Sparse networks from scratch: Faster training without losing performance. *arXiv:1907.04840*, 2019.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[10] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

[13] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv:2106.13230*, 2021.

[14] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CVPR*, 2022.

[15] Mateusz Malinowski, Grzegorz Swirszcz, Joao Carreira, and Viorica Patraucean. Sideways: Depth-parallel training of video models. In *CVPR*, 2020.

[16] Mateusz Malinowski, Dimitrios Vytiniotis, Grzegorz Swirszcz, Viorica Patraucean, and Joao Carreira. Gradient forward-propagation for large-scale temporal video modelling. In *CVPR*, 2021.

[17] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv:1710.03740*, 2017.

[18] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[19] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *NeurIPS*, 2014.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019.

[21] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021.

[22] Chao-Yuan Wu, Ross Girshick, Kaiming He, Christoph Feichtenhofer, and Philipp Krahenbuhl. A multigrid method for efficiently training video models. In *CVPR*, 2020.

[23] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will add instructions in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4's first paragraph and Section 4.1's first paragraph.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We will include these error bars in the supplemental material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4's first paragraph.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited the all datasets we used, such as ImageNet [18] ad COCO [11].
   (b) Did you mention the license of the assets? [No] ImageNet is under BSD 3-Clause License and COCO is under Creative Commons Attribution 4.0 License.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] We do not have sufficient resource to exhaustively screen offensive content from the datasets we used. We expect, because the datasets only contain images of people with daily activities and common objects, they are unlikely to contain offensive content. They are both very widely used.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]