A DATASETS

A.1 LANGUAGE TASKS

MNLI (Multi-Genre Natural Language Inference) corpus (Williams et al., 2018; Wang et al., 2018) is a collection of sentence pairs with textual entailment annotations gathered via crowd sourcing. The sentences are paired as premise and hypothesis and the task is to predict if the premise entails the hypothesis (*entailment*), contradicts the hypothesis (*contradiction*) or neither (*neutral*). The corpus is modeled on the SNLI (Stanford Natural Lanfuage Inference) corpus (Bowman et al., 2015), but differs in that covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation (Williams et al., 2018; Wang et al., 2018). The premises are gathered from ten different sources, including fiction, government reports and transcribed speeches. It consists of 393k train samples and 20k test samples.

QNLI (Question-answering Natural Language Inference) corpus is a dataset automatically derived from SQuAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016; Wang et al., 2018). SQuAD is a question-answering dataset which consists of question-paragraph pairs, where a sentence in the paragraph contains the answer to the corresponding question. QNLI is constructed by converting the task into sentence pair classification by forming a pair between each question and each sentence in the corresponding context. It consists of 105k training samples and 5.4k testing samples.

RTE RTE The Recognizing Textual Entailment (RTE) datasets is a combination of several datasets which came from a series of annual textual entailment challenges Wang et al. (2018); Dagan et al. (2005); Haim et al. (2006); Giampiccolo et al. (2007); Bentivogli et al. (2009). It consists of 2.5k training samples and 3k testing samples.

A.2 VISION TASKS

CIFAR-10 & CIFAR-100 The CIFAR-10 dataset contains 60000 32x32 colour images divided into 10 classes, each with 6000 images. There are 50,000 training and 10,000 test images.

The CIFAR-100 dataset is just like the CIFAR-10, except it has 100 classes containing 600 images each. There are 500 training images and 100 testing images per class. The 100 classes in the CIFAR-100 are grouped into 20 superclasses. Each image comes with a ”fine” label (the class to which it belongs) and a ”coarse” label (the superclass to which it belongs).

The ILSVRC 2012 image classification dataset contains 1.2 million images for training and 50,000 for validation from 1000 classes. The input image sizes are 224×224 center crop to images at test time. The results are reported on the validation set.

B NETWORKS

B.1 ROBERTA

The RoBERTa model is kept the same as its original form proposed by Liu et al. (2019). In our experiment, we consider the RoBERTa-base model only. The base model contains 12 layers, with a hidden size of 768, an FFN inner hidden size of 3072 and 12 attention heads. The original model uses Dropout and we replace all of the original Dropouts to structured Dropout methods.

B.1.1 RESNET

We consider both the original ResNet (He et al., 2016) and its wider alternative (WideResNet) (Zagoruyko & Komodakis, 2016). These networks normally have one convolutional layer (named stem) and four other residual blocks. For the CIFAR10 and CIFAR100 classification, we change the striding of the first convolution to 1 and deleted the first max pooling. These adaptions help the network to operate with the 32×32 image size on CIFAR datasets

B.1.2 PVT-V2

Table 6 demonstrates the detailed setup of the PVT-V2 structure used for CIFAR and ImageNet tasks. The rest of the setup parameters are the same as Wang et al. (2022), and the setup is the same as the PVTV2-B1 model.

Table 6: PVT-V2 setup for vision datasets, e is the embedding dimension and s is the striding used for the overlapping patch embedding.

Layer name	CIFAR10/CIFAR1000	ImageNet
Stage 1	$e = 16, s = 4$	$e = 64, s = 4$
Stage 2	$e = 32, s = 2$	$e = 128, s = 2$
Stage 3	$e = 64, s = 1$	$e = 256, s = 2$
Stage 2	$e = 128, s = 2$	$e = 512, s = 2$

C HARDWARE SYSTEM

We used a variety of hardware systems, our initial testing and CIFAR10 results are generated on a hardware system with 4 x NVIDIA GeForce RTX 2080 Ti GPUs. The ImageNet training and RoBERTa training are performed on 4 x Nvidia A100 SXM4 80GB GPUs. The total amount of GPU training cost for all the expriments in this paper is around 20 GPU-days.

D PICKING THE DROPPING PROBABILITY

Table 7: Different Dropping probabilities for ProbDropBlock.

Probability	ResNet50 on CIFAR10	PVT-V2 on CIFAR100	RoBERTa on MNLI
0.0	94.37 ± 0.32	82.38 ± 0.19	87.60 ± 0.04
0.1	94.70 ± 0.14	$\mathbf{82.44 \pm 0.16}$	$\mathbf{87.83 \pm 0.15}$
0.2	$\mathbf{94.73 \pm 0.19}$	82.21 ± 0.13	87.32 ± 0.11
0.3	94.20 ± 0.30	82.10 ± 0.16	69.45 ± 0.16

E PICKING THE BLOCK SIZE

The block size for DropBlock is a hyperparameter that needs to be tuned. We test $B \in \{2, 4, 6, 8, 10\}$, and pick the best performing B ($B = 4$ in this case).

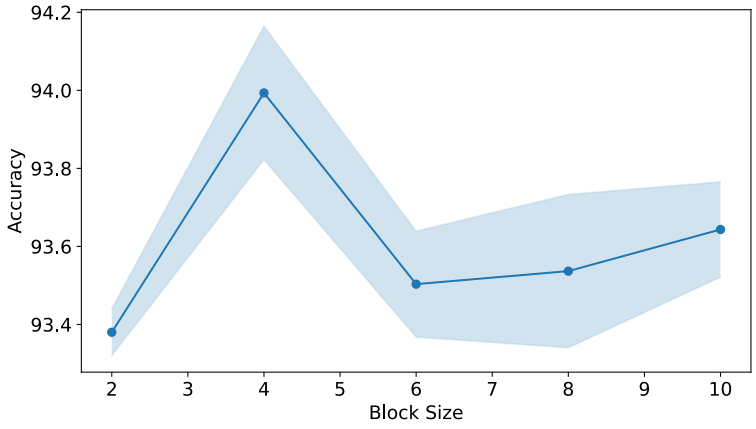

Figure 2: The effect of block size on the performance of DropBlock.