Exploring Task-Agnostic Token-Pruning for ViT with Knowledge Distillation

Ming Y. Lu
MIT EECS
mingylu@mit.edu

Haley Nakamura
MIT EECS
halnak@mit.edu

Abstract

The Vision Transformer (ViT) is a popular emerging architecture, rivaling convolutional neural networks (CNNs) in a variety of important computer vision tasks. However, ViTs are expensive for both training and inference, with inference cost scaling quadratically with the number of image tokens. Expanding on prior research in ViT token-pruning, this paper explores the combination of inattentive token-pruning with knowledge distillation from a task-agnostic self-supervised teacher in order to observe the potential savings in both efficiency and accuracy. Several student-teacher model pairs, token keep rates, and non-distilled models are compared for analysis. Models are evaluated on two lightweight benchmark image classification datasets. Token-pruning 50% of tokens leads to a 50% reduction in MACs and a 2x increase in throughput for a given model, with 0-2.5% decrease in accuracy compared to counterparts which were not pruned. In classification accuracy, knowledge-distilled models outperformed non-pretrained models. Token pruning seems to be more effective for images with a greater background or scenic focus, and more repeated tokens.

1 Introduction

The Vision Transformer[2] (ViT) has become a popular architecture for solving a wide range of computer vision tasks such as image classification, object detection, and semantic segmentation. It is also used extensively in large-scale multimodal and unimodal pre-training for transfer to downstream tasks. Unlike Convolutional Neural Networks (CNNs), the core operation in ViT, multihead self-attention (MHSA), treats each image as a set (sequence) of image tokens (comprised of patches of the image). Its computational complexity scales quadratically with the sequence length (*i.e.* the number of image tokens)[**?**]. The other core operation in ViT applies a non-linear transformation independently to each token using a shared multi-layer perceptron (MLP), which also scales directly with the number of tokens and can readily adapt to inputs of variable sequence lengths without modification. As a result, token pruning has emerged as an intuitive and effective framework for reducing the computational complexity of ViT inference by simply removing "unimportant" tokens from the sequence, without requiring special hardware support.

Another method of network reduction, and also therein inference time reduction, is knowledge distillation. By transferring knowledge from a teacher to a student model, useful information and input embeddings can be distilled into a smaller model which would otherwise require further training, or which may not learn more complex tasks. A common practice for selecting teacher models is using a self-supervised pre-trained model. These models acquire knowledge and valuable embeddings from a large corpus of unlabeled data, and are readily distilled due to their task-agnostic setup. This makes them easily transferable to a variety of downstream tasks, leading to all-purpose input embeddings.

As such, this paper explores the combination of knowledge distillation from a task-agnostic self-supervised teacher model with inattentive token pruning in ViTs. In expansion of previous works, a

6.5940 Course Project.

variety of knowledge distillation settings are tested, including distilling the teacher's global [CLS] token or additionally distilling retained token representations. Similarly, multiple ViT settings with varying ratios of inattentive token pruning are tested. Additional ViT models are also trained from scratch in order to provide baseline comparisons. Models are evaluated for network size and throughput, and performance is tested on two lightweight image classification benchmark datasets.

2 Related Works

ViTs[3] emerged as an exploration into leveraging the success of the long-term relationships and attention of the general transformer architecture[7]. Images are split up into non-overlapping patches which are treated as tokens, and are subsequently projected to a token embedding and sequentially-stacked through the transformer network. In the original ViT architecture, which focuses on image classification, an additional [CLS] token is appended to the patch tokens and is used for classification. The value of the [CLS] token is computed using the attention mechanism across all of the image patch tokens. Dosovitskiy et al., 2021[2] showed that the ViT architecture can match or exceed the accuracy state-of-the-art CNN architectures in multiple image tasks, including image classification. Both Dosovitskiy and Caron et al., 2021[1] additionally show that the attention of the [CLS] token is primarily focused on class-specific objects rather than "background" regions. ViTs additionally display weaker image-specific inductive biases compared to CNNs, strengthening the robustness of the model [1].

EViT[1] introduces a simple method of improving the computational cost of the ViT architecture. As compared to removing tokens with low attention values to the [CLS] token, Liang et al., 2022[3] show that complete removal can weaken the classification accuracy of ViTs. Therefore, EViT merges less attentive tokens into a single token via a weighted average fusion according to their attention value. The total number of tokens in the inference can be adjusted using a token keeping rate, from which the top-k attentive tokens are kept and the remaining tokens are fused. Since the fusion step is a simple linear operation, the cost is negligible compared to the ViT inference cost per token. Using EViT, he authors show significant improvements to inference speed (for example, DeiT-S, EViT achieves a 50% increase in inference speed with a 0.3% decrease in classification accuracy on ImageNet). Compared to some alternative methods proposed in the literature for accelerating transformers, EViT is simple to implement, targets ViT tasks directly, and introduces efficient operations to selectively remove path tokens.

3 Experimental Setup

3.1 Task-Agnostic Knowledge Distillation

First, a series of student models were distilled from SOTA DINOV2[5][2] pre-trained ViT-L vision foundation model, used as the teacher model. All student models were distilled in two different settings: 1) distilling just the global [CLS] token from the teacher model, and 2) distilling both the global [CLS] token along with the representations of the retained patch tokens. We considered three student models: **1) ViT-S**: a small ViT architecture with no token-pruning, **2) EViT-S 0.7**: a small EViT architecture with a token keep rate of 0.7 and **3) EViT-S 0.5**: a small EViT architecture with a token keep rate of 0.5.

Therefore, six different models were distilled (two different settings for each student model).

The MSCOCO[4] training split was used for knowledge distillation, with more information on this dataset in Section 3.5. Weights for EViT-S and ViT-S students were randomly initialized before distillation, and trained for 20 epochs using a batch size of 64 per GPU on 2 NVIDIA 3090 GPUs. We used AdamW and a cosine learning rate scheduler with an initial learning rate of 1e-4.

[1] https://github.com/youweiliang/evit

[2] https://github.com/facebookresearch/dinov2

Model	Params (M)	MACs (M)	Throughput (imgs/s)
ViT-L	304.4	81043.8	72.8
ViT-S	21.6	6127.2	940.9
EViT-S/0.7	21.6	3963.7	1443.1
EViT-S/0.5	21.6	3013.7	1831.3

Table 1: Summary of base model architectures. Throughput is tested with a NVIDIA 3090 GPU, batch size of 32, image size of 224 x 224 in fp32.

3.2 Additional Models for Comparison

Along with the knowledge-distilled models, each of the three student models was also randomly initialized for testing without any pre-training. Additionally, the SOTA DINOv2 model was tested for comparison.

Two additional experiments tested an EViT-S 0.7 model and an EViT-S 0.5 model with weights initialized from the best downstream performing model, the VIT-S knowledge distilled from both the gloabl [CLS] token and the token representations from the DINOv2 teacher. After initialization, the EViT-S models were evaluated on the same downstream tasks outlined in the following section.

3.3 Metrics

For performance, each pre-trained or distilled model will be evaluated using top1 fine-tune multi-class classification accuracy, where accuracy is evaluated on the validation set (or test set, for datasets without a validation split) after every fine-tune epoch and the maximum accuracy is reported. Top1 accuracy is used to align with common benchmarking practices, particularly for the reported results for the benchmark datasets used. Top5 accuracy was not considered since the small amount of classes in the lightweight downstream datasets gives top5 little useful meaning. For each model trained from scratch, top1 validation (or test) accuracy was reported on the relevant split during training from scratch.

For efficiency and size, each network was evaluated in the number of parameters, MACs, and throughput (in images per second). These statistics are reported for each base model architecture in Table 1.

3.4 Downstream Evaluation

Each model described in the previous sections was evaluated on two lightweight image classification benchmarks:

1. Imagenette: a subset of ImageNet with 10 object categories
2. Intel Image Classification: outdoor scenic images with 6 categories

The datasets are further described in Section 3.5 below. Each model was fitted with a linear classification head with the size of the number of object cateogries for the relevant dataset. All distilled models were evaluating in two settings: 1) full fine-tune, in which all parameters are trainable, and 2) linear fine-tune, where all parameters are frozen except for the classification head. The DINOv2 teacher model was only evaluated using linear fine-tune due to the size of the model and restrictions on available computation. Similarly, the models trained from scratch are only reported in full training. All fine-tuning and training methods are performed for 10 epochs with a learning rate of 1e-4.

3.5 Datasets

The datasets below are chosen for their task-relevance and breadth of images, but are lightweight for efficient and reasonable analysis.

Model	Pretrain	Acc (unfrozen)	Acc (frozen)
ViT-S	None	**67.3**	/
EViT-S/0.7	None	66.9	/
EViT-S/0.5	None	66.0	/
ViT-L	DinoV2	/	99.7
ViT-S	KD, CLS	78.8	**76.5**
EViT-S/0.7	KD, CLS	78.2	75.7
EViT-S/0.5	KD, CLS	**79.4**	75.7
ViT-S	KD, All	81.7	**82.4**
EViT-S/0.7	KD, All	81.8	81.7
EViT-S/0.5	KD, All	**81.9**	80.0

Table 2: Performance results of each model on the Imagenette benchmark dataset reported in top1 multi-class classification accuracy. Unfrozen accuracy refers to training or fine-tuning all available parameters. Frozen accuracy refers to training or fine-tuning only the classification head.

Unsupervised Knowledge Distillation:
MSCOCO[3]: 118k images consisting of various scenes and object categories.

Downstream Supervised Evaluation:
ImageNette[4]: a subset of ImageNet with 10 easily distinguishable categories (tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, parachute). There are 13,394 total images (9,469 train and 3,925 validation in the provided split). Imagenette v2 full-size was used for this paper.
Intel Image Classification[5]: a scenic dataset consisting of 25,000 images of size 150×150. There are 6 scene categories (buildings, forest, glacier, mountain, sea, street).

All images were transformed using Torchvision's classification training or evaluation transformation preset[6] (for each use respectively) with a crop size of 224×224.

3.6 Computational Resources

Pretraining was completed using NVIDIA GeForce RTX 3090 GPUs. Downstream evaluation was completed on Google Colab Pro using an NVIDIA T4 Tensor Core GPU.

4 Results

4.1 Model Size and Efficiency

Model summary statistics are provided in Table 1. Here, we observe that while the number of parameters in the small setting for both ViT and EViT is the same, token pruning leads to an approximately proportional reduction in MACs. For example, with a 50% token reduction, MACs are reduced from 6127.2 M in ViT-S to 3013.7 M in EViT-S with keep rate 0.5. Similarly, proportional speedups are acquired in throughput. EViT-S achieves almost double the throughput compared to ViT-S. These findings are consistent with Tang et al.[6], which similarly finds these theoretical and empirical relationships in token reduction stemming from the relationship outlined in Section 1.

4.2 Classification Performance

Classification performance is reported on the Imagenette dataset in Table 2 and the Intel Image Classification dataset in 3. The results of the additional experiments outlined in Section 3.2.

[3] https://cocodataset.org/#home
[4] github.com/fastai/imagenette
[5] kaggle.com/datasets/puneet6060/intel-image-classification/
[6] https://github.com/pytorch/vision/blob/main/references/classification/presets.py#L6-L44

Model	Pretrain	Acc (unfrozen)	Acc (frozen)
ViT-S	None	**78.2**	/
EViT-S/0.7	None	77.6	/
EViT-S/0.5	None	77.8	/
ViT-L	DinoV2	/	94.6
ViT-S	KD, CLS	**85.7**	**84.4**
EViT-S/0.7	KD, CLS	84.6	83.9
EViT-S/0.5	KD, CLS	85.3	83.8
ViT-S	KD, All	**87.1**	**87.3**
EViT-S/0.7	KD, All	86.2	86.5
EViT-S/0.5	KD, All	86.3	86.0

Table 3: Performance results of each model on the Intel Image Classification benchmark dataset reported in top1 multi-class classification accuracy. Unfrozen accuracy refers to training or fine-tuning all available parameters. Frozen accuracy refers to training or fine-tuning only the classification head.

Results from the Imagenette benchmark show several key findings. Firstly, the accuracy for all pre-trained and knowledge distilled models was significantly higher than the equivalent models trained from scratch. This implies that the knowledge distillation transferred effectively to this downstream task, with distilled models achieving upward of 10% higher than counterparts. Similarly, models which were distilled from both the teacher's global [CLS] token and retained token representations performed consistently higher than models which were only distilled from the global [CLS] token. However, margins were tighter, with only a 2-3% increase per model.

As is consistent with the general relationship of fine-tuning, tuning all layers (without freezing) achieved a higher accuracy than only tuning the linear layer, for accuracy gains of approximately 2-4% for an equivalent model. ViT-S linear fine-tuning did outperform the full fine-tune equivalent, but within a small margin.

On Imagenette, linear fine-tuning consistently led to the unpruned ViT-S model outperforming pruned EViT-S models. Contrastingly, EViT-S with a token keep rate of 0.5 outperformed EViT-S 0.7 and ViT-S in full fine-tuning. However, the margin of performance was close, with only a 0.6% gain over the next highest model for [CLS] token distilled models, and a 0.1% gain over the next highest model for fully distilled models.

On the Intel Image Classification dataset, many of the same patters were observed, as is shown in Table 3. Particularly, the relationship of further distilling and further fine-tuning independently lead to equivalently higher results. However, it is also observed that ViT-S consistently outperformed EViT-S for both token keep rates across all experimental settings. Additionally, all models performed better on the Intel Image Classification dataset than their counterparts on the Imagenette dataset, except for the DINOv2 teacher model.

The additional results are provided in Table 4.

4.3 Visualization of Token-Pruning

We find the models can already meaningfully identify tokens from foreground objects even in the absense of task-specific finetuning. Suggesting that the the task-agnostic KD pretraining phase may already be sufficient in helping the models learn which tokens to prune.

5 Discussion

All models performed better on the Intel Image Classification dataset than the Imagenette dataset, except for the DINOv2 teacher model. One potential factor in this is the distribution of token patches in the input images. The Intel Image Classification images are scenic and do not generally contain foreground objects. They are often dominated by sky or land, with many repeated tokens of landscape colors. In contrast, Imagenette contains images of foreground of objects which require more unique

Model	Pretrain	Acc (unfrozen)	Acc (frozen)
ImageNette			
EViT-S/0.7	KD, All	81.8	**81.7**
EViT-S/0.5	KD, All	**81.9**	80.0
EViT-S/0.7	ViT-S KD, All	81.4	80.6
EViT-S/0.5	ViT-S KD, All	81.7	76.8
Intel Image Classification			
EViT-S/0.7	KD, All	86.2	**86.5**
EViT-S/0.5	KD, All	86.3	86.0
EViT-S/0.7	ViT-S KD, All	**86.9**	85.7
EViT-S/0.5	ViT-S KD, All	78.9	86.3

Table 4: Performance results from additional experimentation. We investigate whether there is a benefit to using a full ViT-S pretrained using the KD, All configuration and only performing token pruning in downstream tasks. We did not see a clear benefit to this approach suggesting it may be important to expose the model to token pruning early on in the task-agnostic phase.

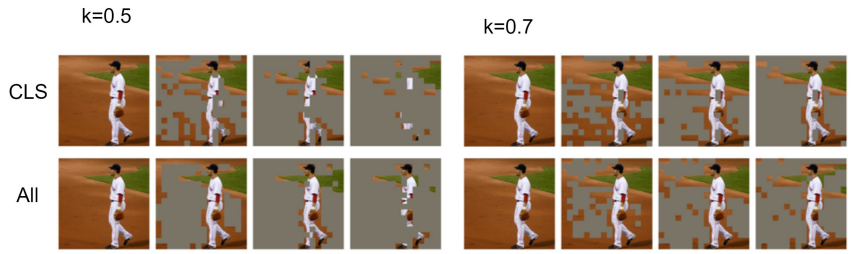

Figure 1: Visualization of token pruning, example 1. The pretrained EViT-S/0.5 and EViT-S/0.7 are used (before any task-specific finetuning).

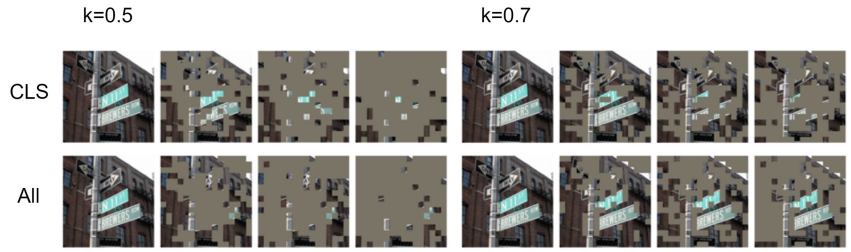

Figure 2: Visualization of token pruning, example 2. The pretrained EViT-S/0.5 and EViT-S/0.7 are used (before any task-specific finetuning).

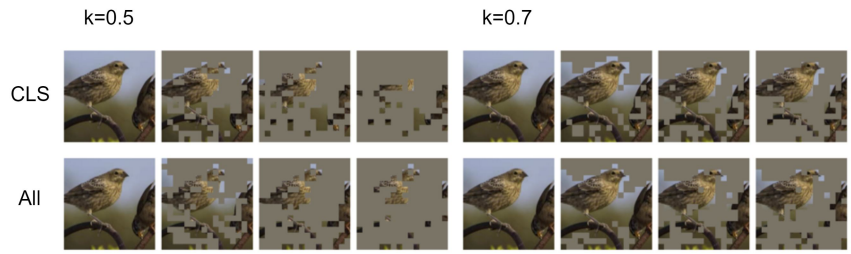

Figure 3: Visualization of token pruning, example 3. The pretrained EViT-S/0.5 and EViT-S/0.7 are used (before any task-specific finetuning).

Figure 4: Visualization of token pruning, example 4. The pretrained EViT-S/0.5 and EViT-S/0.7 are used (before any task-specific finetuning). We find in the absense of a single foreground object, the models retains tokens from diverse objects in the scene.

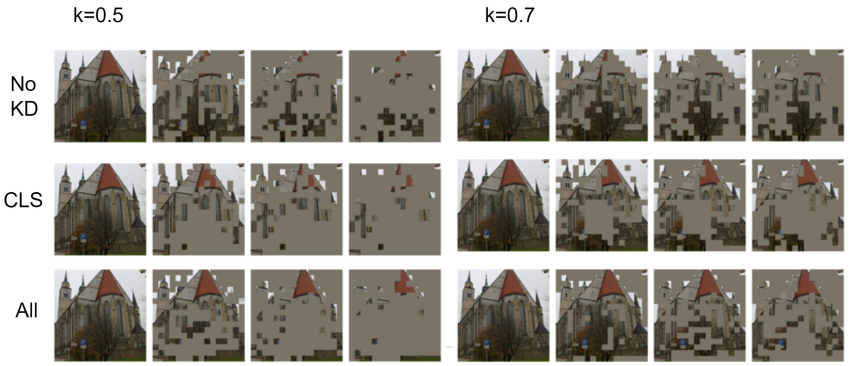

Figure 5: Visualization of token pruning after finetuning on ImageNette.

tokens that contain edges, object features, and more. This would make token pruning generally more conducive to scenic imagery, for which many tokens are already repetitive. Generalized to all ViTs, the general patterns of the scenic images are also more distinct, and there are fewer categories to learn, which suggests that lightweight models could more easily distinguish classes regardless of token pruning.

In general, both ViT-S and EViT-S transferred effectively to downstream image classification. Results showed that further knowledge distillation and further fine-tuning lead to better performance. Aggressive token pruning (for example, the 0.5 keep rate case) led to an approximately 50% reduction in MACs and a doubling in throughput. However, in different experimental settings, a gap in performance remains for EViT compared to equivalently trained ViT-S, ranging from between 0%-2.5%. In settings where EViT performed better, the margin of improvement was minimal.

Future work could explore additional token keep rates to observe the effects in downstream accuracy. Additionally, it would be prudent to transfer these experiments to other downstream tasks, such as image segmentation or depth estimation, in order to understand the relationship between pruning and the features relevant to other tasks.

6 Conclusion

This work evaluated the performance of models across several different settings of combined token pruning and knowledge distillation from a large self-supervised model. Experiments showed that knowledge distillation of more representations increased accuracy for downstream image classification, most exacerbated when compared to models trained from scratch. Token pruning led to proportional decrease in MACs and increase in throughput, with comparatively smaller losses in accuracy. Token pruning visualizations show the proportion of tokens dropped, with background patches being pruned first (across different token keep rates). Token pruning displayed a strong positive relationship with knowledge distillation in regards to network reduction, inference speedup, and accuracy retention.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[2] 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 preprint arXiv:2010.11929*, 2020.

[3] Y Liang. Evit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations.*, 2022.

[4] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[6] Yehui Tang, Kai Han, Yunhe Wang, Chang Xu, Jianyuan Guo, Chao Xu, and Dacheng Tao. Patch slimming for efficient vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.