Bootstrapping ViTs: Towards Liberating Vision Transformers from Pre-training

Haofei Zhang1,*, Jiarui Duan1,*, Mengqi Xue1, Jie Song1,†, Li Sun1, Mingli Song1,2
1Zhejiang University 2 Shanghai Institute for Advanced Study of Zhejiang University

Abstract

Recently, vision Transformers (ViTs) are developing rapidly and starting to challenge the domination of convolutional neural networks (CNNs) in the realm of computer vision (CV). With the general-purpose Transformer architecture replacing the hard-coded inductive biases of convolution, ViTs have surpassed CNNs, especially in data-sufficient circumstances. However, ViTs are prone to over-fit on small datasets and thus rely on large-scale pre-training, which expends enormous time. In this paper, we strive to liberate ViTs from pre-training by introducing CNNs’ inductive biases back to ViTs while preserving their network architectures for higher upper bound and setting up more suitable optimization objectives. To begin with, an agent CNN is designed based on the given ViT with inductive biases. Then a bootstrapping training algorithm is proposed to jointly optimize the agent and ViT with weight sharing, during which the ViT learns inductive biases from the intermediate features of the agent. Extensive experiments on CIFAR-10/100 and ImageNet-1k with limited training data have shown encouraging results that the inductive biases help ViTs converge significantly faster and outperform conventional CNNs with even fewer parameters. Our code is publicly available at https://github.com/zhfeing/Bootstrapping-ViTs-pytorch.

1. Introduction

The great successes of the convolutional neural networks (CNNs) [21, 23, 29, 42] have liberated researchers from handcrafting visual features [14, 33]. By means of the inductive biases [10], i.e., focusing on the localized features and weight sharing, CNNs are potent tools for tackling visual recognition tasks [5, 21, 36]. Nevertheless, such biases have constrained their abilities towards building deeper and larger models, as they have ignored the long-range dependencies [15, 18].

In recent years, Transformers [46] have been proposed for replacing inductive biases with a general-purpose network architecture in natural language processing (NLP). Exclusively relying on multi-head attention mechanisms (MHA), Transformers have the inborn capability to capture the global dependencies within language tokens and have become the de facto preferred data-driven models in NLP [2, 17, 35]. Inspired by this, a growing number of researchers have introduced the Transformer architecture into the realm of computer vision (CV) [4, 18, 45, 56]. It turns out an encouraging discovery that vision Transformers (ViTs) outperform state-of-the-art (SOTA) CNNs by a large margin with a similar amount of parameters.

Despite the appealing achievements, ViTs suffer from poor performance, especially without adequate annotations or strong data augmentation strategies [8, 18, 45]. The reasons for this circumstance are two fold: on the one hand, the widely adopted multi-head self-attention mechanisms (MHSA) in ViTs have dense connections against convolution [11], which is hard to optimize without prior knowledge; on the other hand, Chen et al. [8] have illustrated that ViTs tend to converge to minima with sharp regions, usually related to limited generalization capability and overfitting problems [7, 25]. Therefore, the typical training scheme of Transformers in NLP [2, 17] relies on the large-scale pre-training and then fine-tuning for downstream tasks, which consume enormous GPU (TPU) time and energy [4, 18, 45]. For instance, Dosovitskiy et al. [18] spend thousands of TPU days to pre-train a ViT with 303M images. Spontaneously, it raises the following question: how can we optimize ViTs efficiently without pre-training.
To the best of our knowledge, existing approaches focused on the problem can be mainly divided into two parts. The first line of approaches attempts to bring inductive biases back into Transformers, such as sparse attention [6, 12, 27] and token aggregation [52]. Such heuristic modifications to ViTs will inevitably lead to the sophisticated tuning of plenty of hyperparameters. The second line of approaches [8, 24, 45] aims at constructing suitable training schemes for Transformers, which helps them converge with better generalization ability. In particular, Chen et al. [8] utilize the sharpness-aware minimizer (SAM) [20] to find smooth minima, while [24, 45] optimize a Transformer by distilling knowledge from a pre-trained teacher. Nonetheless, these methods still require pre-training on mid-sized datasets, such as ImageNet-1k [29], and how to efficiently train ViTs with relatively small datasets from scratch remains an open question.

Motivated by the distillation approaches [22, 24, 45] that utilize a teacher model to guide the optimization direction of the student, in this paper, we strive to make one step further towards optimizing ViTs with the help of an agent CNN, which also learns from scratch along with the ViT. Our goal is to inject the inductive biases from the agent CNN into the ViT without modifying its architecture and design a more friendly optimization process so that ViTs can be customized on small-scale datasets without pre-training.

To this end, we propose a novel optimization strategy for training vision Transformers in the bootstrapping form so that even without pre-training on mid-sized datasets or strong data augmentations, ViTs can still be competitive when lack of training data. Specifically, as shown in Fig. 1, we first propose an agent CNN designed corresponding to the given ViT, and with the inductive biases, the agent will converge faster than the ViT. Then we jointly optimize the ViT along with the agent in the mutual learning framework [55], where the intermediate features of the agent supervise the ViT with the inductive biases for fast convergence. In order to reduce the training burden, we further share the parameters of the ViT to the agent and propose a bootstrapping learning algorithm to update the shared parameters. We have conducted extensive experiments on CIFAR-10/100 datasets [28] and ImageNet-1K [29] under the lack-of-data settings. Experimental results demonstrate that: (1) our method has successfully injected the inductive biases to ViTs as they converge significantly faster than training from scratch and eventually surpass both the agents and SOTA CNNs; (2) the bootstrapping learning method can efficiently optimize the shared weights without the extra set of parameters.

Our contributions are summarized as three folds:

1. We propose agent CNNs constructed based on standard ViTs, for training ViTs efficiently with shared weights and inductive biases.
2. We propose a novel bootstrapping optimization algorithm to optimize the shared parameters.
3. Our experiments show that ViTs can outperform the SOTA CNNs even without pre-training by adopting both inductive biases and suitable optimizing goals.

2. Related Work

2.1. Vision Transformers

With the powerful self-attention mechanisms, the Transformer [46] has been the SOTA and preferred model in NLP [2, 3, 17]. Inspired by the impressive success of Transformers in NLP, researchers are starting to introduce Transformers for tackling CV tasks. ViT [18] is a groundbreaking work that utilizes the pure Transformer architecture for image classification and has achieved great successes. The variants of the ViT [9, 31, 45, 48, 52, 57] are further utilized for more complex CV tasks, e.g., semantic segmentation [40, 56] and object detection [4, 13]. However, ViTs rely on large-scale pre-training and have shown poor performance with limited training data. To address this problem, some approaches try to introduce inductive biases to ViTs with heuristic modification, e.g., sparse attention [6, 12, 27], token aggregation [52]. The others aim to propose novel training schemes tailored for Transformers [8, 24, 45]. Nonetheless, these methods still require pre-training on mid-scale datasets, such as ImageNet-1K. It still remains an open question of how to optimize ViTs efficiently without pre-training, especially on small-scale datasets. To resolve this problem, we strive to inject CNNs’ inductive biases into ViTs without modifications to the network architecture.

2.2. Knowledge Distillation

Knowledge distillation (KD) [22] is an effective model compression technique which the hidden knowledge of the teacher is transferred to the student by supervising with soft labels. To sufficiently transfer the knowledge, FitNets [37] additionally use intermediate features for supervision, and the following works [39, 44, 53] extract the deeper level of information in different aspects. More recently, mutual learning [55], a variant of KD, has attracted many interests as all the models (students) are learning from each other simultaneously. This practical learning strategy has been applied to person re-identification [19, 47], object detection [49], and face recognition [1, 16]. Apart from the applications of mutual learning, some researchers focus on improving mutual learning by introducing more supervision like the intermediate features [50] or feature fusion [26]. Inspired by this, we propose utilizing an agent CNN optimized jointly with the ViT. The hard-coded inductive biases are transferred to the ViT under the mutual learning framework with adaptive intermediate feature supervision.
3. Method

In this section, we first introduce the preliminaries of CNNs and ViTs. Then, based on the relationship of convolution layers and MHSA layers, agent CNNs are proposed to help train ViTs. Finally, we delineate the bootstrapping optimization algorithm where the agent and ViT are jointly optimized without pre-training. The workflow of our method is demonstrated in Fig. 2.

3.1. Preliminaries

3.1.1 Convolution

Convolution is the central part of CNNs, which accepts a two-dimensional feature map. For the sake of future discussion, we formalize the forward propagation with a sequence of visual tokens $X = (x_1, \ldots, x_n) \in \mathbb{R}^{n \times d_{in}}$ as its input, each of which is a $d_{in}$-dimensional embedded vector. Thus, the output sequence of convolution with kernel size $(k_h, k_w)$ is the sum of linear projection of $X$:

$$Y_C = \sum_{i=1}^{N} \Phi_i X W_i,$$  \hfill (1)

where $\Phi_i$ is a constant sparse matrix representing the hard-coded inductive biases of localized dependencies, the size of receptive field $N = k_h \times k_w$, and the projection matrix $W_i \in \mathbb{R}^{d_{in} \times d_{out}}$ is trainable$^1$.

It is worth noting that the $1 \times 1$ convolution has the form of $Y = X W$, which is equivalent to a fully connected (FC) layer with the same projection matrix $W$.

---

$^1$For simplicity, the bias terms are omitted. The detailed derivations are presented in Appendix A.

3.1.2 MHSA

The multi-head self-attention mechanism (MHSA) in ViTs takes a sequence of visual tokens as its input and can also be formularized similar to Eq. (1):

$$Y_M = \sum_{h=1}^{H} \Psi_h X W_{VO}^h,$$  \hfill (2)

where $H$ is the number of heads, $W_{VO}^h = W^V W^O$ is the combination of two projection matrices ($W^V_h \in \mathbb{R}^{d \times d_k}$, $W^O_h \in \mathbb{R}^{d_k \times d}$, $d = H d_k$), and $\Psi_h \in \mathbb{R}^{n \times n}$ is the dense attention matrix computed based on the pair-wise similarity of linearly projected tokens.

3.2. Agent CNN

Inspired by the similarity of Eq. (1) and Eq. (2) that the convolution layer can be treated as a special case of MHSA layer with sparse relationship matrices $\Psi$, we propose constructing an agent CNN based on a given ViT that will converge faster when trained from scratch.

3.2.1 Generalized Convolution

To begin with, we propose a generalized convolution layer in which the size of its receptive field $N$ equals to the head number $H$ of a MHSA layer, named as CONV, with hard-coded inductive biases $\{\Phi_h\}_{h=1}^{H}$:

$$Y_{CONV} = \sum_{h=1}^{H} \Phi_h X W_h,$$  \hfill (3)
where \( \Phi_1, \ldots, \Phi_H \in \mathbb{R}^{n \times n} \) are extracted from the set of hard-coded inductive biases \( \{ \Phi_1, \ldots, \Phi_N \} \) of the \( \sqrt{H} \times \sqrt{H} \) convolution \( (N' = \sqrt{H}^2) \) defined in Eq. (1).

### 3.2.2 Constructing Agent CNN

We start with a standard ViT model (in Fig. 3a) with \( m \) encoder layers and finally build an agent CNN for introducing the inductive biases of CNNs. Fig. 3b illustrates a base agent CNN by simply replacing the MHSA layers of ViT with CONV layers, which has introduced sparsity and localized biases. Besides, the MLP in the agent are composed of two \( 1 \times 1 \) convolution layers which are equivalent to the fully connected layers in the FFNs of Transformers.

Moreover, as many preferred CNNs share the feature pyramid architecture \([21, 30, 38, 54]\) that the spacial size of feature maps shrunken as going deeper, we construct the final res-like agent CNN (in Fig. 3c) by: (1) introducing a ResNet-style input projection block which contains two convolution layers and one max-pooling layer, (2) adopting a configurable down-sample after each encoder layer.

With the hard-coded inductive biases, the agent can converge faster and with higher performance than training the corresponding ViT from scratch as shown in Fig. 5c.

### Weight Sharing

Utilizing the homologous network architecture, our proposed agent accepts shared weights from the ViT model to reduce the training burden. Due to the equivalence of \( 1 \times 1 \) convolution and FC layers, the FFNs in each encoder block of ViT can be directly shared by the agent. Furthermore, when shared with the output projection \( W^{VO} \) of MHSA in Eq. (2), the CONV has the form of

\[
Y_{\text{CONV}} = \sum_{h=1}^{H} \Phi_h X W^{VO}_h.
\]

Let \( y_c \) and \( \tilde{y}_c \) be the \( c \)-th token of the output of MHSA and shared CONV accordingly. With the assumption that the input sequences are the same, denoted by \( X \), the difference \( y_{\text{err}} = y_c - \tilde{y}_c \) can be written as

\[
y_{\text{err}} = \sum_{h=1}^{H} (\bar{\phi}_h - \psi_h) X W^{VO}_h,
\]

where \( \bar{\phi}_h \) and \( \psi_h \) are the \( c \)-th row of matrix \( \bar{\Phi}_h \) and \( \Psi_h \) respectively. As there are no more than one non-zero element in \( \bar{\phi}_h \) (proved in Appendix A), we can minimize the magnitude of \( y_{\text{err}} \) by learning sparse and localized dependencies of matrix \( \Psi_h \).

### 3.3. Bootstrapping Optimization

In this section, we describe how to jointly optimize the ViT and the agent by introducing the optimization objective and proposed training strategy.

#### 3.3.1 Adaptive Intermediate Supervision

To inject the inductive biases of the agent into ViT without modifying ViT’s architecture, we propose the adaptive intermediate supervision where the adapted feature maps of the agent supervise the corresponding visual sequences of the ViT. Let \( F_A^\ell \) and \( F_V^\ell \) denote the flattened feature maps and visual sequence of the \( \ell \)-th encoder layer of the agent and ViT in respective. The adaptive intermediate loss of \( \ell \)-th layer of the ViT and agent is defined as

\[
\mathcal{L}_{\text{feat}}^{\ell} = \left\| \frac{\bar{F}_A^\ell (\Phi_A^\ell)}{\| F_A^\ell (\Phi_A^\ell) \|_2} - \frac{F_V^\ell}{\| F_V^\ell \|_2} \right\|_2^2.
\]

where \( \bar{F} := \text{Adapt}(F) \) is the adapted feature, obtained from sequence interpolation or 2-dimensional average pooling. We have compared the different adaptive approaches in Sec. 4.3.

Finally, the adaptive intermediate supervision is the sum of all assigned layers \( A \):

\[
\mathcal{L}_{\text{feat}} = \sum_{\ell \in A} \mathcal{L}_{\text{feat}}^{\ell}.
\]

### 3.3.2 Optimization Objective

Except for intermediate supervision, we introduce the mutual learning framework \([55]\) that the predicted probabilities
by the ViT (denoted as $p_V$) and the agent (denoted as $p_A$) learn from each other as

$$L_{\text{mutual}} = L_{\text{KD}}(p_V, \bar{p}_A, y; T) + L_{\text{KD}}(p_A, \bar{p}_V, y; T),$$

where $\bar{p}$ represents that the variable $p$ is treated as a constant vector, i.e., no gradient is computed with regard to the variables in the forward propagation paths, $L_{\text{KD}}$ is the knowledge distillation loss defined in [22] with temperature $T$, and $y$ denotes the ground truth label of the input image.

Above all, the optimization objective is summarized as

$$L = 3, \quad \beta L_{\text{mutual}},$$

and $\alpha$ and $\beta$ are the weighting hyperparameters for balancing the two terms.

### 3.3.3 Bootstrapping Training Algorithm

The bootstrapping training algorithm is given in Algorithm 1 where gradients computed from each network are aligned and jointly update the shared weights. The gradient alignment function $\text{Align}(\nabla_S, \nabla_P)$ modifies the negative gradient direction from the agent as [51].

**Relationship with multi-task learning.** It is worth pointing out that bootstrapping learning is different from multi-task learning. As shown in Fig. 4, the multi-task model $E$ only accepts one input $X$, while for bootstrapping learning, inputs to $E$ and $E'$ are different. Moreover, the layers $E$ and $E'$ in Fig. 4a shares the same weights $\Theta$. In our case, as we constrain the difference between inputs of each encoder layer by Eq. (7) and when $\|y_{\text{valid}}\|$ is small enough, the bootstrapping learning will degenerate to multi-task learning.

### 4. Experiments

#### 4.1. Implementation

**Datasets.** Three widely-used image classification datasets are adopted to evaluate our proposed method as benchmarks, including CIFAR-10 [28], CIFAR-100 [28], and

![Algorithm 1 Bootstrapping optimizer for training the shared weights in FFN and MHSA layers in the ViT.](#)

**Input:** $\Theta_S$: the set of shared weights; $\Theta_V$, $\Theta_A$: the set of private weights in the ViT and agent respectively; $E_V^{(l)}(:; \Theta_S, \Theta_V)$, $E_A^{(l)}(:; \Theta_S, \Theta_A)$: the $l$-th encoder layer of the ViT and agent respectively; $\lambda$: learning rate.

1. **while not converge do**
2. Compute the input feature map (sequence) $X_V, X_A$ to the $l$-th encoder layer from the input image
3. $Y_V \leftarrow E_V^{(l)}(X_V; \Theta_S, \Theta_V)$
4. $Y_A \leftarrow E_A^{(l)}(X_A; \Theta_S, \Theta_A)$
5. Compute the gradients $\nabla_Y^V$, $\nabla_Y^A$ w.r.t. $\Theta_S$ and $\Theta_V$ respectively ($\Theta_A$ keep as constant variables).
6. Compute the gradients $\nabla_S$, $\nabla_P$ w.r.t. $\Theta_S$ and $\Theta_A$ respectively ($\Theta_V$ keep as constant variables).
7. $\Theta_V \leftarrow \Theta_V - \lambda \nabla_Y^V$, $\Theta_A \leftarrow \Theta_A - \lambda \nabla_P$ # update private weights
8. $\Theta_S \leftarrow \Theta_S - \lambda \text{Align}(\nabla_S, \nabla_P)$ # update shared weights in bootstrapping way
9. **end while**

ImageNet-1k [29]. In particular, to simulate lack-of-data circumstances, 1%, 5%, and 10% labeled samples are randomly extracted from the training partition of the ImageNet dataset. Even though ViTs require strong data augmentations in previous approaches [18, 45, 48, 52], in our implementation, both CNNs and ViTs are only optimized with several simple augmentation methods, including random resized cropping and random horizontal flipping.

**Vision Transformers.** We follow the network architectures introduced in [18] and [45] with slight modifications to the head number and embedding dimensions. The detailed settings of ViTs are presented in Tab. 2, in which ViT-S is a relatively small model with 6 layers, and the ViT-B is the same as DeiT [45] with 12 layers.

**Agent CNNs.** The agents are constructed according to the given ViTs and thus share the same network settings as ViTs. Besides, detailed configurations of the base and residualized agents are described in Appendix B.

**Training Details and Selection of Hyperparameters.** We implement our method with Pytorch [34] framework. AdamW [32] is used to optimize ViTs and agent CNNs in both standalone and joint training schemes, with the learning rate of $10^{-3}$ and the weight decay of $5 \times 10^{-2}$. However, conventional CNNs, such as ResNet [21] and EfficientNet [43], are optimized by SGD [41] with the learning rate of $5 \times 10^{-2}$ and the weight decay of $5 \times 10^{-4}$. We train all the settings for 240 epochs with the batch size of
Table 1. Comparison results on CIFAR-10 and CIFAR-100. The top-1 accuracy, number of parameters, and FLOPs are reported separately. ‘†’ indicates that the initial weights of pre-trained ViT-B are acquired from the official repository of DeiT. The comparison settings are classified in the ‘Model’ column. The values in blue color indicates top-1 accuracy improvements compared with the corresponding ViT trained from scratch.

<table>
<thead>
<tr>
<th>Model</th>
<th>Method</th>
<th>CIFAR-10</th>
<th></th>
<th>CIFAR-100</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc</td>
<td>#param.</td>
<td>FLOPs</td>
<td>Acc</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CNNs</td>
<td>EfficientNet-B2</td>
<td>94.14</td>
<td>7.71M</td>
<td>0.70G</td>
<td>75.55</td>
</tr>
<tr>
<td></td>
<td>ResNet50</td>
<td>94.92</td>
<td>23.53M</td>
<td>4.14G</td>
<td>77.57</td>
</tr>
<tr>
<td></td>
<td>Agent-S</td>
<td>94.18</td>
<td>8.66M</td>
<td>3.37G</td>
<td>74.62</td>
</tr>
<tr>
<td></td>
<td>Agent-B</td>
<td>94.83</td>
<td>25.05M</td>
<td>9.46G</td>
<td>74.78</td>
</tr>
<tr>
<td>ViTs</td>
<td>ViT-S</td>
<td>87.32</td>
<td>6.28M</td>
<td>1.37G</td>
<td>61.25</td>
</tr>
<tr>
<td></td>
<td>ViT-S-SAM</td>
<td>87.77</td>
<td>6.28M</td>
<td>1.37G</td>
<td>62.60</td>
</tr>
<tr>
<td></td>
<td>ViT-S-Sparse</td>
<td>87.43</td>
<td>6.28M</td>
<td>1.37G</td>
<td>62.29</td>
</tr>
<tr>
<td></td>
<td>ViT-B</td>
<td>79.24</td>
<td>21.67M</td>
<td>4.62G</td>
<td>53.07</td>
</tr>
<tr>
<td></td>
<td>ViT-B-SAM</td>
<td>86.57</td>
<td>21.67M</td>
<td>4.62G</td>
<td>58.18</td>
</tr>
<tr>
<td></td>
<td>ViT-B-Sparse</td>
<td>83.87</td>
<td>21.67M</td>
<td>4.62G</td>
<td>57.22</td>
</tr>
<tr>
<td>Pre-trained</td>
<td>ViT-S</td>
<td>95.70</td>
<td>6.28M</td>
<td>1.37G</td>
<td>80.91</td>
</tr>
<tr>
<td>ViTs</td>
<td>ViT-B†</td>
<td>97.17</td>
<td>21.67M</td>
<td>4.62G</td>
<td>84.95</td>
</tr>
<tr>
<td>Ours Joint</td>
<td>Agent-S</td>
<td>94.90</td>
<td>8.66M</td>
<td>3.37G</td>
<td>74.06</td>
</tr>
<tr>
<td></td>
<td>ViT-S</td>
<td>95.14 (+7.82)</td>
<td>6.28M</td>
<td>1.37G</td>
<td>76.19 (+14.94)</td>
</tr>
<tr>
<td></td>
<td>Agent-B</td>
<td>95.06</td>
<td>25.05M</td>
<td>9.46G</td>
<td>76.57</td>
</tr>
<tr>
<td></td>
<td>ViT-B</td>
<td>95.00 (+15.76)</td>
<td>21.67M</td>
<td>4.62G</td>
<td>77.83 (+24.76)</td>
</tr>
<tr>
<td>Ours Shared</td>
<td>Agent-S</td>
<td>93.22</td>
<td>-</td>
<td>-</td>
<td>74.06</td>
</tr>
<tr>
<td></td>
<td>ViT-S</td>
<td>93.72 (+6.40)</td>
<td>6.28M</td>
<td>1.37G</td>
<td>75.50 (+14.25)</td>
</tr>
<tr>
<td></td>
<td>Agent-B</td>
<td>92.66</td>
<td>-</td>
<td>-</td>
<td>74.11</td>
</tr>
<tr>
<td></td>
<td>ViT-B</td>
<td>93.34 (+14.10)</td>
<td>21.67M</td>
<td>4.62G</td>
<td>75.71 (+22.64)</td>
</tr>
</tbody>
</table>

Table 2. Detailed configurations of ViTs and agent CNNs.

<table>
<thead>
<tr>
<th>ViT-S</th>
<th>ViT-B</th>
<th>Agent-S</th>
<th>Agent-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layers (m)</td>
<td>6</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Hidden Size (d)</td>
<td>288</td>
<td>384</td>
<td>288</td>
</tr>
<tr>
<td>Heads (H)</td>
<td>9</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

32 on two Nvidia Tesla A100 GPUs. The cosine annealing is adopted as the learning rate decay schedule.

There are several hyperparameters involved in our method, including $\alpha$ and $\beta$ in Eq. (9) and the temperature $T$ for knowledge distillation loss in Eq. (8). We set $\alpha = 1$, $\beta = 10$ and $T = 4$ as default values and the sensitivity analysis is described in Sec. 4.5.

Finally, we set the intermediate feature supervisions decay linearly for preserving the capacity of the ViTs, and different decay strategies are compared in Sec. 4.3.

4.2. Experimental Results

We evaluate our proposed method with the following comparison settings:

- **CNNs**: the standalone agents and traditional CNNs, e.g., EfficientNet-B2 and ResNet50.
- **ViTs**: the original ViTs and variants for training with high efficiency, e.g., ViT-Sparse [12] and ViT-SAM [8]. All of them are trained from scratch.
- **Pre-trained ViTs**: vision Transformers which have been pre-trained on ImageNet-1k and then fine-tuned to evaluation datasets.
- **Ours Joint**: the vision Transformers and the agent are jointly optimized without weight sharing.
- **Ours Shared**: the vision Transformers and the agent are jointly optimized with weight sharing.

**Performance on CIFAR Datasets.** The comparison results on CIFAR-10/100 are shown in Tab. 1, where the top-1 accuracy, number of parameters, and FLOPs are presented for each setting. The discoveries are itemized as follows. (1) When the agent CNNs are trained alone, the performance can hardly surpass the traditional CNNs, even if...
most of the inductive biases are hard-coded into the agents. In Sec. 4.4, we discuss more about the choice of agents. (2) Without pre-training or strong data augmentations, the ViTs perform terribly due to the dense connection of MHSA layers. Although Chen et al. [8] have shown that ViTs can be optimized with SAM optimizer, inherited from the pre-training on mid-scale datasets, such as ImageNet-1k, it may not be the best choice for small datasets like CIFAR. (3) Our proposed method has significantly outperformed the baseline settings, including the original ViTs and the variations. Particularly, ViT-S surpasses the original baseline by 7.82% on CIFAR-10 and 14.49% on CIFAR-100, which outperforms the agent and EfficientNet-B2 with fewer parameters. Besides, both the ViTs and agents have benefited from our proposed method; however, with the help of the global receptive field, ViTs perform better. (4) In the ‘Shared’ settings, the bootstrapping learning strategy has shown that the shared weights can be optimized robustly with a limited deterioration of accuracy. Such results are encouraging that the weights of ViTs can be directly transferred to a framework of hard-coded inductive biases so that ViTs can utilize the inductive biases without an extra set of parameters or sophisticated modifications.

Moreover, we plot the accuracy learning curves of our method along with the baseline settings in Fig. 5. It clearly illustrates that ViTs can converge as fast as CNNs and finally achieve the higher upper bound than CNNs’.

### Performance on ImageNet

The comparison results on ImageNet-1k with different amounts of labeled images are shown in Tab. 3, where 5%, 10%, and 50% of the training images are randomly selected. The conclusions of the CIFAR-10/100 dataset still hold in the ImageNet. Particularly, the improvement of our method is prominent when data is extremely scarce, while others have shown inconspicuous amelioration or even impairment.
Table 5. Joint training results on CIFAR-100 when using the agent with different network architectures.

<table>
<thead>
<tr>
<th></th>
<th>Acc</th>
<th>Agent-S</th>
<th>ViT-S-Joint</th>
<th>Agent-S-Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>72.73</td>
<td>73.18</td>
<td>73.79</td>
<td></td>
</tr>
<tr>
<td>Res-like</td>
<td>74.78</td>
<td>76.19</td>
<td>74.06</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Ablation Study

In this section, we go deep into our proposed method to figure out the function of loss terms (introduced in Sec. 3.3.2), the effects of different decay strategies (in Sec. 4.1), and adaptive functions (in Sec. 3.3.1). The results of ViT-S are reported on CIFAR-100 dataset.

4.3.1 Ablation of Loss Terms

As delineated in Sec. 3.3.2, the ultimate optimization objective has two terms: the adaptive intermediate supervision $L_{\text{feat}}$ and the mutual learning term $L_{\text{mutual}}$. We evaluate the joint learning settings when supervised with the two terms separately in Tab. 4, showing that both loss terms have contributed to the final result. Particularly, $L_{\text{feat}}$ boosts the accuracy by 13.69 percentage points. Additionally, learning curves are plotted in Fig. 5c for better illustration. We can observe that the ‘Feat’ term produces a more competitive result as it converges significantly faster than using the ‘Mutual’ term. Therefore, the supervision through the intermediate features has successfully injected the inductive biases into the ViT.

4.3.2 Ablation of Decay Strategy

The influence of the feature supervision decay strategy is shown in Tab. 4. Without the decay strategy, the performance has declined by 1.35%. It can be explained that the constant supervision with inductive biases has constrained the ViTs from learning the long-range dependencies, and consequently, impaired the upper-bound of ViTs.

4.3.3 Ablation of Adaptive Functions

We evaluate our method with two intermediate feature adaptive functions: 1D sequence interpolation (by default) and 2D average pooling. The comparison results are shown in Tab. 4, in which the sequence interpolation is superior to the average pooling.

4.4. The Choice of Agent CNNs

In Sec. 3.2.2, we have introduced agent CNNs with two different network architectures (base and res-like). Tab. 5 shows that the performances with the res-like configuration are uniformly better than the base agent.

4.5. Sensitivity Analysis of Hyperparameters

The sensitivity analysis of the hyperparameters in our methods is illustrated in Fig. 6, including $\alpha$, $\beta$ in Eq. (9), and $T$ in Eq. (8). It shows that our method is robust to the variation of $\alpha$ and $T$. However, $\beta$ has a more significant influence, as the ViTs perform better when more inductive biases are used to supervise the ViTs.

5. Conclusion and Future Work

In this paper, we propose to dissolve optimizing vision Transformers (ViTs) efficiently without pre-training or strong data augmentations. Our goal is to introduce inductive biases from convolutional neural networks (CNNs) to ViTs while preserving the network architecture of ViTs for higher upper bound, and further setting up more suitable optimization objectives. To this end, we propose to optimize the ViT jointly with an agent CNN constructed corresponding to the network architecture of the ViT. The ViT learns inductive biases through adaptive intermediate supervision and predicted probabilities. Besides, a bootstrapping training algorithm is proposed to optimize both the ViT and agent with weight sharing. Extensive experiments have shown encouraging results that the inductive biases help ViTs converge significantly faster and outperform conventional CNNs with fewer parameters. In future work, we will extend our method beyond CNN-style inductive biases and introduce more interpretable features to ViTs.

Acknowledgements. This work is supported by Key Research and Development Program of Zhejiang Province (2020C01024), CCF-Baidu Open Fund (NO.2021PP15002000), National Natural Science Foundation of China (62106220, U20B2066), Ningbo Natural Science Foundation (2021J189), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (Grant No. SN-ZJU-SIAS-001), and the Fundamental Research Funds for the Central Universities.
References


