2022「中技社科技獎學金」
2022 CTCI Foundation Science and Technology Scholarship
境外生研究獎學金
Research Scholarship for International Graduate Students

Establishing a novel learning method to efficiently train Vision Transformers on tiny datasets

Jules Rostand, 2nd year PhD student, advisors : Professor Chen-chien James Hsu & Professor Cheng-Kai Lu
Department of Electrical Engineering, National Taiwan Normal University

1. Abstract

While transformer models are becoming more and more relevant in the world of computer vision, they perform poorly when trained with limited data, on which they are still outperformed by their CNN counterparts. **Li & al.** recently have shown that SOTA pure vision transformer's can be improved on tiny datasets by using CNN locality guidance [1], in plug and play fashion. Despite considerable improvements, on top of ResNet counterparts still showing higher accuracy results, the guidance adds slight overhead, further increasing the gap in training time between CNNs and VTs of similar sizes. To enhance this method, I have so far altered the learning method to considerably improve the accuracy (**+6.1%** on CIFAR-100 and **+2.2%** onChaoyang dataset when using DeiT architecture) while reducing the added training time induced by the guidance.

2. Research Focus

The concept of locality guidance is mainly based on the observation that it is hard for a transformer architecture to learn with limited data due to the high flexibility and intrinsic globality of the mechanism that is at its core, the self-attention. On the other hand, convolutional architectures are able to benefit from their inductive biases, such as translational invariance and locality, to outperform VTs of similar sizes on tiny datasets. Therefore, **Li & al.** [1] first train a lightweight CNN network using a very small resolution (32*32 pixels) which takes very little time, and then train the VT architecture by adding an element to the loss function:

$$L = L_{cls} + \beta L_{guidance},$$

Where **Lcls** is the usual loss obtained via supervised classification, **Lguidance** is a guidance loss based on the difference between the weights of the neurons of the CNN and that of the VT, and β is the weight value associated to the guidance loss. After resizing each layer's outputs for the VT and the CNN's sizes to match, the guidance loss is calculated in this fashion :

$$L_{guidance} = \sum_{i=1}^{k} \frac{1}{H_i * W_i} * \left\| F_{vt}^{i} - F_{cnn}^{i} \right\|_F^2 \Big|$$

, where **i** lists the different layers of the network.

That way, on top of still benefitting from the VT's ability to learn the global context, the network also captures the local information using the CNN guidance.

We make the hypothesis that, similarly to the way fine-tuning from a pre-trained model works, we actually need **early** guidance, more than guidance. We further conjecture that while the CNN guidance enhances the learning of the VT in the early stages of the VT, it restrains it in the latter stages. To verify this, we progressively decrease the weight associated to the guidance loss during the first 100 epochs and discount the guidance loss during the next 200 epochs.

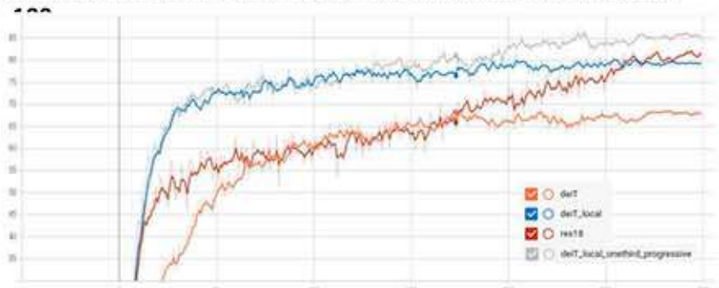
3. Datasets

We first experiment on the **CIFAR-100** dataset (100 classes, 50000 training images and 10000 validation images) and a dataset related to medical imagery named **Chaoyang** dataset consisting of 6160 representing 4 different classes of pathology detection.

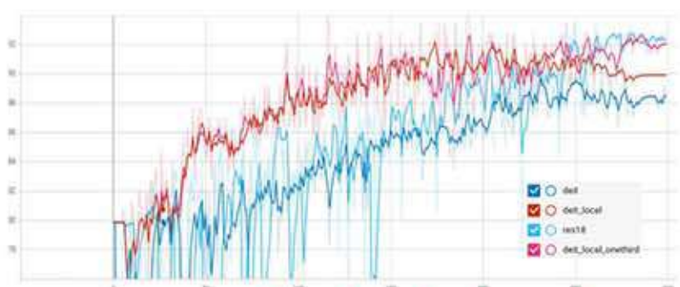
4. Results

We first train our lightweight CNN (**ResNet-56**) used for guidance, then we compare the baseline VT, the VT using local guidance, the VT using local guidance and our learning method, and a CNN of comparable size (**ResNet-18**). In these experiments, we compare training times and Top-1 accuracies using **DeiT-tiny** as our baseline VT.

Evolution of the validation accuracy across epochs on CIFAR-



Evolution of the validation accuracy across epochs on Chaoyang



Training times and accuracies

Model	CIFAR-100		Chaoyang	
	Accuracy(%)	Training time	Accuracy(%)	Training time
DeiT	67.82	4h 45m	88.50	57m
DeiT-local	79.35	5h 17m	89.93	58m
DeiT-local-ours	**85.43**	**4h 58m**	**92.06**	**57m**
ResNet-18	81.66	3h 43m	**92.24**	55m

5. Conclusion and perspectives

By altering the learning method, we not only considerably reduce the added training time induced by the guidance, we also give the VT more freedom to learn by itself in the latter stages of the training, which leads to a massive boost in accuracy, achieving similar results than ResNet-18 on Chaoyang dataset, and passing it by almost 4% on CIFAR-100, which hasn't been done yet with a pure vision transformer. As I continue with my research, I intend to further enhance my approach to obtain a novel learning method, as well as test its efficiency on additional tiny datasets and vision transformer architectures.

6. References

[1] Li, Kehan and Yu, Runyi and Wang, Zhennan and Yuan, Li and Song, Guoli and Chen, Jie : Locality Guidance for Improving Vision Transformers on Tiny Datasets. arXiv preprint arXiv:2207.10026 (2022)