IEEE International Conference on Multimedia and Expo 2023
Brisbane Convention & Exhibition Centre
10-14 July 2023

南京大学 NANJING UNIVERSITY

Preserving Locality in Vision Transformers for Class Incremental Learning

Bowen Zheng[1] Da-Wei Zhou[1] Han-Jia Ye[1] De-Chuan Zhan[1]

[1]National Key Laboratory for Novel Software Technology, Nanjing University

July 12th, 2023

Contents

Class Incremental Learning

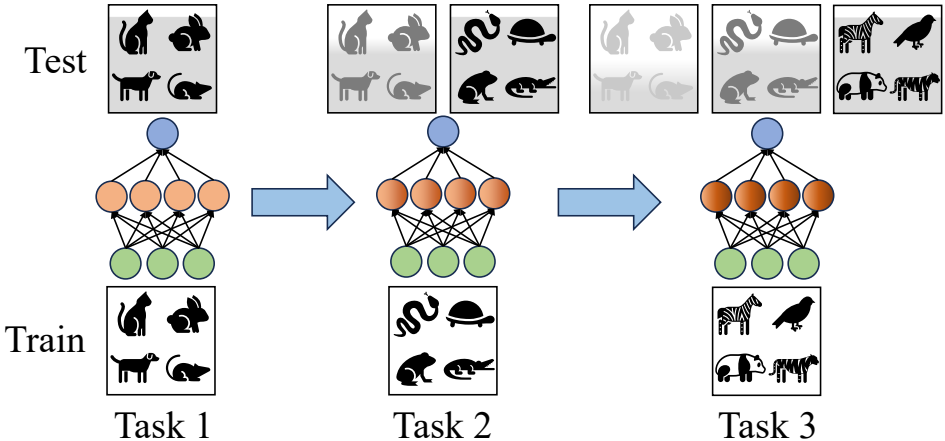

Figure 1: Class Incremental Learning (CIL) and Catastrophic Forgetting

Motivation

- We study an aspect of vision Transformers used in CIL, which is *Locality*.
- Locality means the model's ability to capture local features.
- Locality is not properly preserved in CIL.

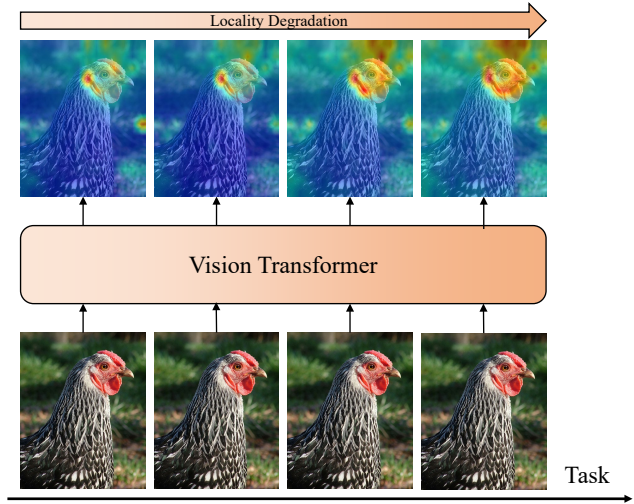

Figure 2: Locality Degradation

Locality Degradation

Attention Heat Map Visualization

- Joint: Joint Learning, where all of the presented tasks are trained together at each stage.
- In CIL, locality degrades as the task goes on, comparing to joint learning.

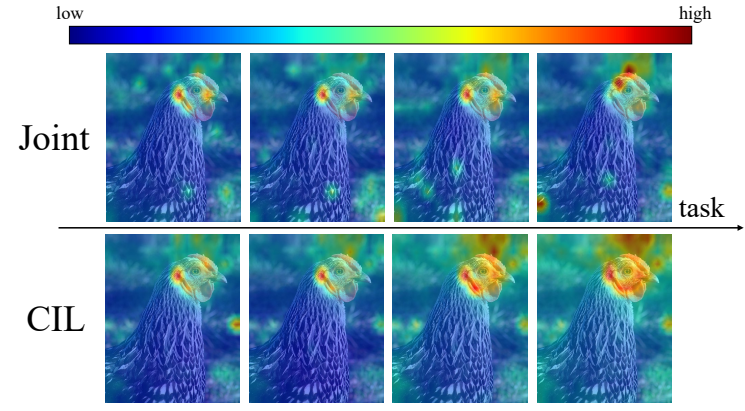

Figure 3: Attention Heat Map Visualization Examples

Locality Degradation

Quantitative Nonlocality Measure

- Nonlocality of a layer:

$$D_{loc}^{(l)(h)} = \frac{1}{N} \sum_{ij} \boldsymbol{A}_{ij}^{(l)(h)} \|\boldsymbol{\delta}_{ij}\|, \quad (1)$$

$$D_{loc}^{(l)} = \frac{1}{N_h} \sum_{h} D_{loc}^{(l)(h)}, \quad (2)$$

- For most layers, especially for deep layers, the nonlocality increases as the task goes on.
- Joint learning always has less nonlocality for each layer.

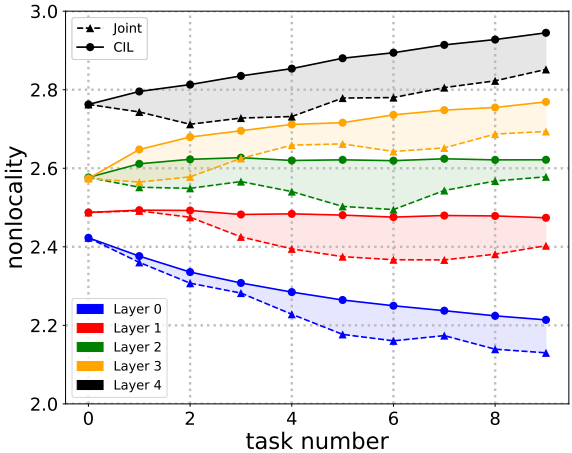

Figure 4: Nonlocality Measure

Why the Difference?

- Shallow layers are more transferable between tasks than deep layers.
- In the prior experiments, shallow layers has more locality.
- The model is losing task-agnostic information during incremental learning.

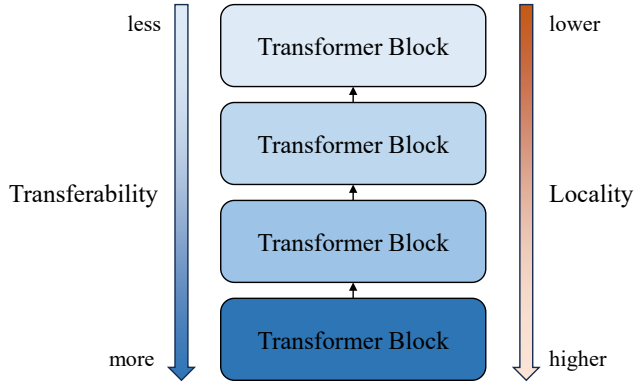

Figure 5: Transferability and Locality

Locality-Preserving Attention (LPA)

- We directly introduce the locality into the unnormalized attention score.
- The attention now has two parts, the global attention score $\boldsymbol{A}^{(h)}$ and the local attention score $\boldsymbol{v}_h^\top \boldsymbol{r}$.
- The new attention score for each head:

$$\tilde{\boldsymbol{A}}^{(h)} = \mathrm{softmax}(\lambda_h \boldsymbol{A}^{(h)} + \boldsymbol{v}_h^\top \boldsymbol{r}), \tag{3}$$

- We assign a weight λ_h to mix them up, and initialize it to a small value to control the gradients at the initial steps of the training.

Experiments

Locality Preserving

- We use LPA to replace each layer of the vision transformer.
- The LPA layer successfully preserves the locality of ViT during CIL, compared to prior locality degradation experiment.

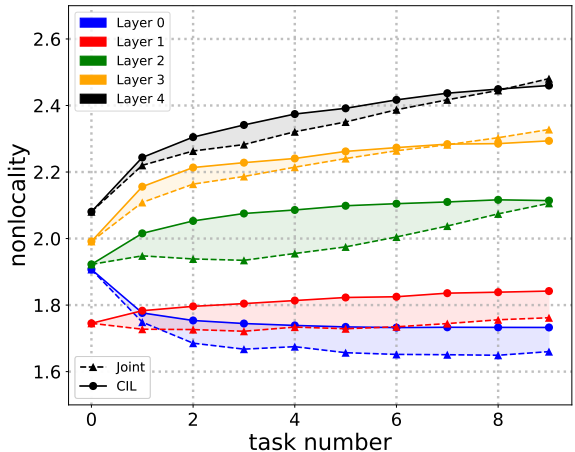

Figure 6: Locality Preserving

Experiments

Performance

Table 1: Performance Results on CIFAR100

Scenarios		baseline	baseline w/LPA	DyTox+	DyTox+ w/LPA
B10-10	Last	63.11	**65.27**	66.79	**68.92**
	Avg	74.74	**76.50**	77.66	**78.74**
	Fgt	12.52	**11.20**	15.36	**14.30**
B5-5	Last	60.23	**60.38**	62.60	**63.99**
	Avg	72.91	**73.14**	76.02	**77.21**
	Fgt	12.47	**12.45**	21.48	**20.73**
B50-10	Last	66.20	**68.52**	69.64	**69.76**
	Avg	73.70	**74.96**	76.00	**76.19**
	Fgt	11.82	**10.02**	**9.464**	9.708
B50-5	Last	63.77	**65.54**	65.70	**66.71**
	Avg	71.86	**73.08**	73.30	**74.27**
	Fgt	14.84	**13.22**	15.13	**13.57**

Table 2: Performance Results on ImageNet100

Scenarios		baseline	baseline w/LPA	DyTox+	DyTox+ w/LPA
B10-10	Last	61.02	**61.98**	65.78	**67.54**
	Avg	72.84	**74.81**	76.35	**77.85**
	Fgt	**14.38**	16.16	17.89	**17.02**
B50-10	Last	68.10	**70.74**	71.32	**71.70**
	Avg	75.62	**76.83**	78.08	**78.46**
	Fgt	15.86	**13.93**	**9.42**	9.78
B50-5	Last	65.4	**67.9**	66.38	**68.08**
	Avg	74.54	**75.38**	75.46	**76.26**
	Fgt	19.75	**18.30**	16.92	**16.31**

Experiments

Representation Transferability

- We also investigate the eigenvalue distribution of representation's covariance matrix.
- With larger eigenvalues, the representation transfer better across tasks.
- With LPA, the eigenvalues are larger, which means there are more transferable directions in the learned representation.

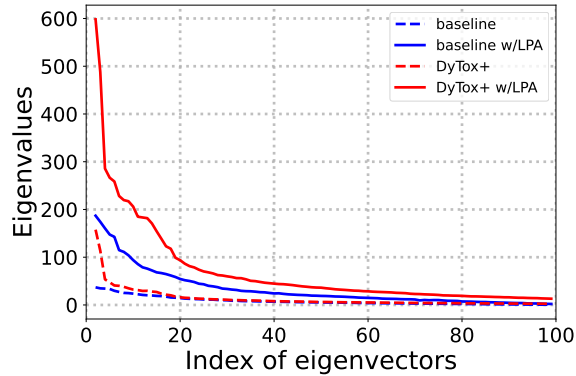

Figure 7: Representation Transferability

Experiments

λ Initialization

- λ is the weight for the global part of the attention.
- The best region is around 0.02.

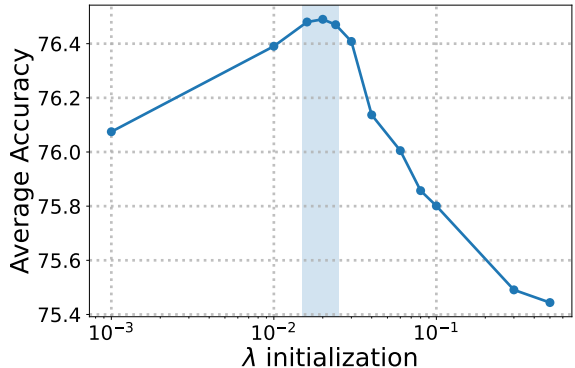

Figure 8: Analysis on λ Initialization

Experiments

Number of LPA Layers

- We replace the attention layers one by one from the shallow layers to deep layers.
- With more locality-preserved attention layers, the performance of incremental learning can be steadily improved.

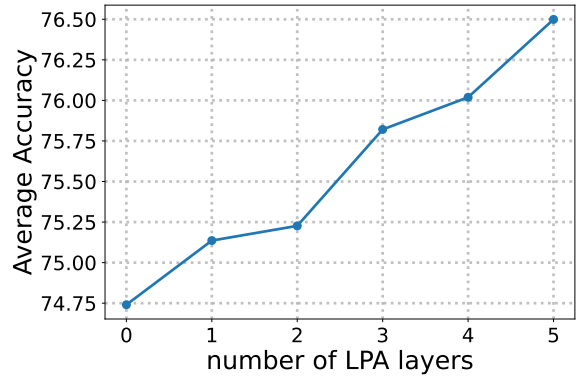

Figure 9: Analysis on Number of LPA Layers

Summary

- We discover the locality degradation in ViTs for CIL, and illustrate it with visualizations and quantitative nonlocality measure.
- We attribute this phenomenon to losing task-agnostic information during CIL.
- We propose LPA to preserve the locality in CIL.
- We performed experiments to verify the preserved locality and improved performance and transferability.

Thank you for listening!