Adapters Strike Back

Jan-Martin O. Steitz[1] and Stefan Roth[1,2]
[1]TU Darmstadt [2]hessian.AI

TECHNISCHE UNIVERSITÄT DARMSTADT

visual inference

SEATTLE, WA CVPR JUNE 17-21, 2024

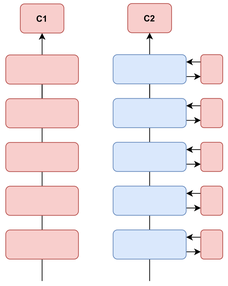
Motivation

Traditional transfer learning

- Need to store all parameters from fine-tuning
- Requires a lot of training data for target task

Adapter-based transfer learning

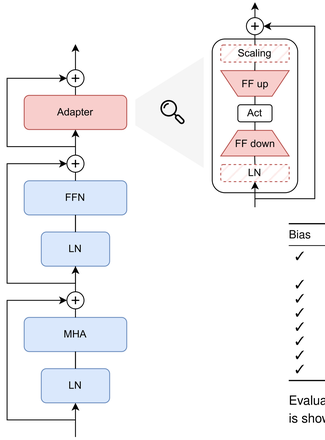
- Fine-tune only adapter modules and classifier
- Very parameter efficient: Only 0.1–0.4% of parameters required compared to fully fine-tuning a ViT-B/16
- Previous implementations introduce issues that lead to their underperformance



TL;DR

We revisit adapters and provide the first in-depth study for their use with vision transformers. We show that adapters can "strike back" and outperform more complex adaptation methods.

Analysis: Inner Adapter Structure



Transformer Layer with Post-Adapter

Which elements are important for the inner structure of the adapter?

We evaluate the effect of

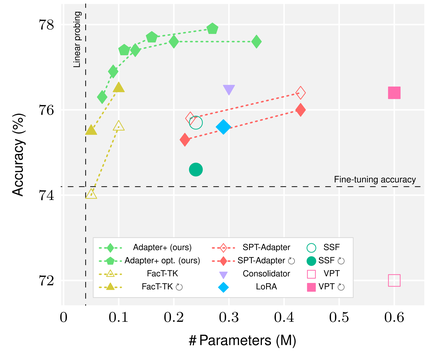
- Biases in linear layers
- Normalization layer
- Learned, layer-wise or channel-wise scaling
- Initialization of the adapter's parameters

Bias	Norm	Scaling	Initialization	Accuracy (%)	Δ_{base}
✓			Houlsby	76.0	0.0
			Houlsby	75.6	−0.4
✓			LoRA	75.5	−0.5
✓			BERT	75.8	−0.2
✓	✓		Houlsby	75.9	−0.1
✓	✓	layer	Houlsby	75.9	−0.1
✓		layer	Houlsby	76.2	+0.2
✓	✓	channel	Houlsby	75.8	−0.2
✓		channel	Houlsby	**76.5**	**+0.5**

Evaluated on the VTAB *val sets*. The difference to Adapter$_{base}$ (first row) is shown in Δ_{base}.

The optimal structure has a channel-wise scaling but no normalization layer and uses Houlsby initialization.

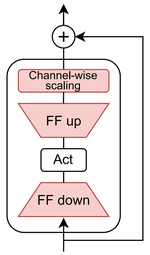
Adapter+



Adapter+ "strikes back": An improved adapter architecture with learned, channel-wise scaling

- Requires little to no manual tuning for adaptation to novel task
- Outperforms previous adapter implementations and more complex adaptation methods
- Reaches state-of-the-art average accuracy on the VTAB and FGVC benchmarks
- Excellent parameter-accuracy trade-off compared to other work

Parameter-accuracy characteristics of adaptation methods on the VTAB *test sets*. ◌: Re-evaluations with suitable data normalization and without early-stopping based on the *test set*.



Inner structure of Adapter+ with channel-wise scaling

Configuration	# Param (M)	Natural	Specialized	Structured	Average
Houlsby, $r=8$	0.39	82.9	85.5	58.9	75.8
Houlsby, $r=4$	0.24	82.9	84.9	58.3	75.4
Pfeiffer	0.21	82.9	86.1	58.4	75.8
AdaptFormer	**0.19**	**83.0**	85.0	57.4	75.2
Adapter+	0.20	**83.0**	**86.8**	**59.7**	**76.5**

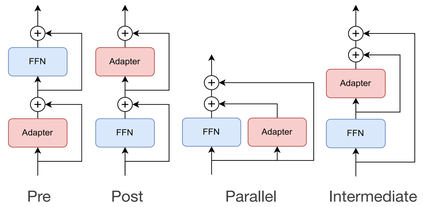
Comparison of Adapter+ with different adapter configurations. Average accuracy (in %) evaluated on the VTAB *val sets*.

Analysis: Adapter Position

We can connect adapters in various ways to the FFN of the transformer layer. Which position is optimal?

Position	Natural	Specialized	Structured	Average
Pre	82.4	**86.2**	57.5	75.3
Intermediate	**83.0**	85.0	57.2	75.1
Parallel	**83.0**	**86.2**	57.7	75.6
Post	**83.0**	85.7	**59.1**	**76.0**

Accuracy (in %) evaluated on the VTAB *val sets*.



The Post-Adapter is the best configuration in the vision transformer layer.

Results

We evaluate Adapter+ on the VTAB and FGVC benchmarks

- Adapter+ sets a new state-of-the-art for parameter-efficient adaptation
- Best on VTAB even without per-task hyperparameter optimization

VTAB

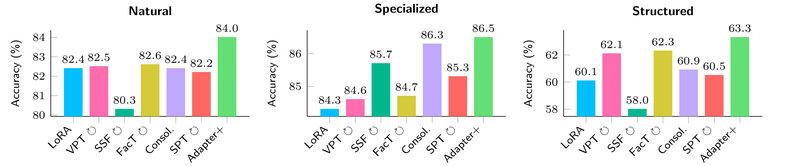
	# Param (M)	Natural: Caltech101	Natural: CIFAR-100	Natural: DTD	Natural: Flower102	Natural: Pets	Natural: SVHN	Natural: Sun397	Natural: Average	Specialized: Camelyon	Specialized: EuroSAT	Specialized: Resisc45	Specialized: Retinopathy	Specialized: Average	Structured: Clevr-Count	Structured: Clevr-Dist.	Structured: DMLab	Structured: KITTI-Dist.	Structured: dSpr-Loc.	Structured: dSpr-Ori.	Structured: sNORB-Azi.	Structured: sNORB-Ele.	Structured: Average	Global Average
Full	85.8	73.2	92.6	70.4	97.9	86.2	90.6	39.6	78.6	87.1	96.6	87.5	74.0	86.3	66.6	61.0	49.8	79.7	82.6	51.9	33.5	37.0	57.8	74.2
Linear	0.04	78.1	88.1	69.0	99.1	90.0	36.0	56.9	73.9	79.8	90.7	73.7	73.7	79.5	32.4	30.5	25.9	61.9	11.2	26.2	14.3	24.5	29.6	61.0
LoRA	0.29	83.0	91.7	71.6	99.2	90.9	83.8	56.7	82.4	86.2	95.7	83.5	71.9	84.3	77.7	62.3	49.0	80.2	82.2	51.7	31.0	47.0	60.1	75.6
VPT-Deep ◌	0.60	83.0	93.0	71.2	99.0	91.3	84.1	56.0	82.5	84.9	96.6	82.5	74.5	84.6	77.5	58.7	49.7	79.6	86.2	56.1	**37.9**	**50.7**	62.1	76.4
NOAH f | ★	0.43	69.6	92.7	70.2	99.1	90.4	86.1	53.7	80.2	84.4	95.4	83.9	75.9	84.9	82.8	**68.9**	49.9	81.7	81.8	48.3	32.8	44.2	61.3	75.5
SSF f ◌	0.24	61.9	92.3	73.4	**99.4**	**92.0**	90.8	52.0	80.3	86.5	95.8	87.5	72.8	85.7	77.4	57.6	53.4	77.0	78.2	54.3	30.3	36.1	58.0	74.6
FacT-TK$_S$ f ◌	**0.05**	74.9	92.7	73.7	99.1	91.3	85.5	57.7	82.1	86.8	94.9	84.1	70.9	84.2	81.9	64.1	49.2	77.2	83.8	53.1	28.2	44.7	60.3	75.5
FacT-TK$_{32}$ f ◌	0.10	74.6	93.7	73.6	99.3	90.6	88.7	57.5	82.6	87.6	95.4	85.5	70.4	84.7	**84.3**	62.6	51.9	79.2	85.5	52.0	36.4	46.6	62.3	76.5
Consolidator ^{2}	0.30	74.2	90.9	**73.9**	**99.4**	91.6	91.5	55.5	82.4	86.9	95.7	86.6	**75.9**	86.3	81.2	68.2	51.6	**83.5**	79.8	52.3	31.9	38.5	60.9	76.5
SPT-Adapter ◌	0.22	74.7	94.1	73.0	99.1	91.2	84.5	57.5	82.0	85.7	94.9	85.7	70.2	84.1	81.3	63.2	49.1	80.7	83.5	52.0	26.4	41.5	59.7	75.3
SPT-Adapter ◌	0.43	74.9	93.2	71.6	99.2	91.1	87.9	57.2	82.2	87.0	95.4	86.5	72.4	85.3	81.1	63.2	50.3	80.2	84.4	51.4	31.5	42.2	60.5	76.0
Adapter+, $r=1$	0.07	**85.4**	92.4	73.1	99.1	91.3	83.1	**58.1**	83.2	87.2	96.6	85.3	72.6	85.5	80.7	60.6	50.9	79.9	83.3	55.6	27.1	43.0	60.1	76.3
Adapter+, $r=2$	0.09	**85.4**	93.0	72.7	99.2	90.6	85.3	56.0	83.5	**87.9**	96.8	85.5	71.4	85.4	83.2	61.0	51.6	80.1	86.1	**56.3**	30.7	46.5	61.9	76.9
Adapter+, $r=4$	0.13	84.8	93.8	72.7	99.2	90.6	86.5	57.4	83.6	87.5	96.9	85.9	71.5	85.4	83.4	61.6	53.6	81.4	87.3	55.3	34.4	48.1	63.1	77.4
Adapter+, $r=8$	0.20	84.6	**84.2**	72.3	99.3	90.7	87.6	56.7	83.6	87.7	**97.0**	86.7	72.3	85.9	83.2	60.9	**53.8**	80.3	88.1	55.8	35.7	47.7	63.1	77.6
Adapter+, $r=16$	0.35	83.7	**84.2**	71.5	99.3	90.6	88.2	55.8	83.3	87.5	**97.0**	87.4	72.9	86.2	82.9	60.9	53.7	80.8	**88.4**	55.2	37.3	46.9	**63.3**	77.6
Adapter+, $r\in[1..4]$ f	0.11	**85.4**	93.8	72.7	99.1	90.6	86.5	**58.1**	83.7	87.5	96.8	85.9	71.4	85.4	83.4	61.0	53.6	81.4	87.3	55.3	34.4	48.1	63.1	77.4
Adapter+, $r\in[1..8]$ f	0.16	**85.4**	93.8	72.7	99.1	90.7	87.6	**58.1**	83.9	87.7	96.8	86.7	72.3	85.9	83.4	60.9	**53.8**	80.3	88.1	55.3	35.7	47.7	63.1	77.7
Adapter+, $r\in[1..32]$ f	0.27	**85.4**	93.8	72.7	99.1	90.7	88.2	**58.1**	**84.0**	87.5	96.8	**87.6**	73.9	**86.5**	83.4	60.9	**53.8**	80.3	87.2	55.3	**37.9**	47.7	**63.3**	**77.9**

Detailed accuracy results (in %) on the VTAB *test sets*. *f*: Per-task hyperparameter optimization, ◌: re-evaluation with suitable data normalization and without early-stopping based on the *test set*.

FGVC

	# Param (M)	CUB200	NABirds	Oxford Flowers	Stanford Dogs	Stanford Cars	Average
Full	86.0	88.0	81.5	99.2	85.6	90.6	89.0
Linear	0.18	88.9	81.8	99.5	92.6	52.8	83.1
VPT-Deep ◌	0.85	**90.1**	83.3	**99.6**	90.3	85.0	89.7
SSF ◌	0.39	88.9	**85.0**	**99.6**	88.9	88.9	90.3
SPT-Adapter	0.40	89.1	83.3	99.2	91.1	86.2	89.8
SPT-LoRA	0.52	88.6	83.4	99.5	91.4	87.3	90.1
Adapter+, $r\in[1..32]$	**0.34**	90.0	83.2	**99.6**	91.6	89.1	**90.7**

Detailed accuracy results (in %) on the FGVC *test sets*. ◌: Re-evaluations with suitable data normalization and without early-stopping based on the *test set*.



Analysis: Regularization and Data Normalization

We investigate the influence of training regularization for transfer learning with adapters. We use Stochastic Depth and additional Dropout inside the adapter.

		Adapter: Stochastic Depth	Adapter: Dropout	Adapter: None
ViT	Stochastic Depth	**76.0**	75.4	75.3
	None	74.5	74.3	73.7

Average accuracy (in %) evaluated on the VTAB *val sets*.

Using a suitable data normalization is essential so as not to lose adaptation capabilities to compensate for the shift in normalization.

	ImageNet norm: Natural	ImageNet norm: Specialized	ImageNet norm: Structured	ImageNet norm: Average	Inception norm: Natural	Inception norm: Specialized	Inception norm: Structured	Inception norm: Average	$\Delta_{Average}$
VPT	79.2	83.0	53.8	72.0	82.2	86.2	57.9	75.4	3.4
LoRA	78.4	84.1	53.2	71.9	82.0	85.8	56.4	74.7	2.8
FacT-TK	78.0	83.3	56.1	72.4	81.6	85.6	58.1	75.1	2.7
Adapter+	**80.5**	**85.0**	**56.0**	**73.9**	**83.0**	**86.8**	**59.7**	**76.5**	**2.6**

Average accuracy (in %) evaluated on the VTAB *val sets*.

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center.

emergenCITY

LOEWE Exzellente Forschung für Hessens Zukunft

Code is available