What Deep CNNs Benefit from Global Covariance Pooling: An Optimization Perspective

Qilong Wang[1], Li Zhang[1], Banggu Wu[1], Dongwei Ren[1], Peihua Li[2], Wangmeng Zuo[3], Qinghua Hu[1]

[1]Tianjin University, [2]Dalian University of Technology, [3]Harbin Institute of Technology

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Paper *Code*

Motivation and Contributions

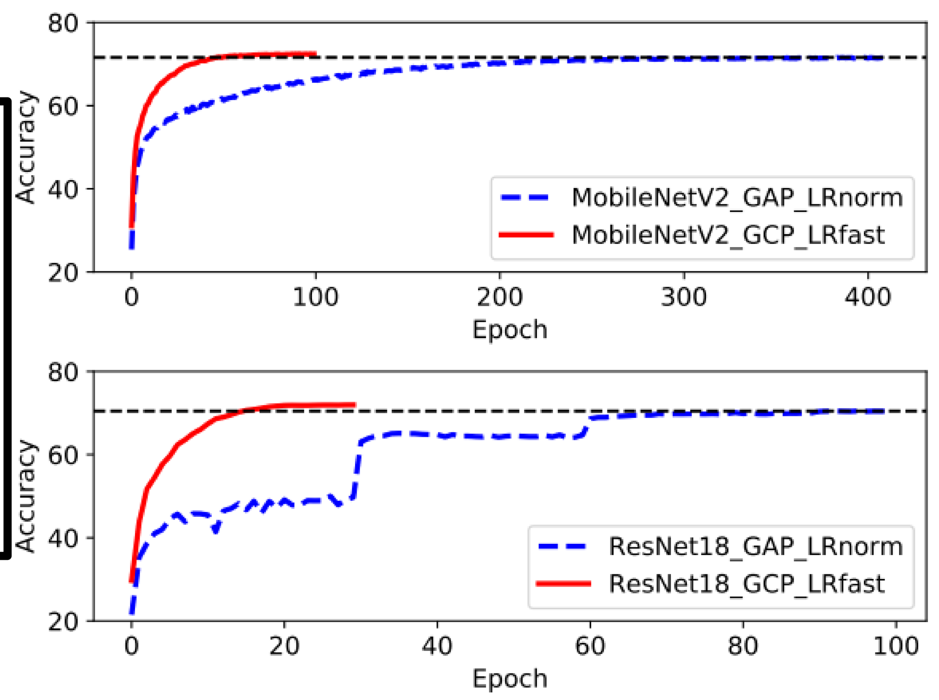
Motivation:

Recent works have demonstrated that global covariance pooling (GCP) has the ability to improve performance of deep CNNs.

- ***Fine-grained Visual Recognition (4~10% gains)***
- ***ImageNet Classification (2~6% gains)***
- ***Texture Classification (~4% gains)***

Despite considerable advance, the reasons on effectiveness of GCP for deep CNNs have not been well studied.

E.g., Why GCP can significantly speed up convergence of deep CNNs?



Contributions:

- ***The first attempt to understand the effectiveness of GCP in the context of deep CNNs from an optimization perspective.***
- ***Showing and explaining several merits of GCP for training deep CNNs that have not been recognized previously or fully explored.***

An Optimization Perspective for GCP

1. Smoothing Effect of GCP

■ ***Definitions[1]:***

- ***Stability of optimization loss (i.e., Lipschitzness):***

$$\Delta_l = \mathcal{L}(\mathbf{X} + \eta_l \nabla_{\mathbf{X}} \mathcal{L}(\mathbf{X})),\ \eta_l \in [a, b]$$

- ***Stability of gradients (i.e., predictiveness):***

$$\Delta_g = \left\| \nabla\mathcal{L}(\mathbf{X}) - \mathcal{L}\left(\mathbf{X} + \eta_g \nabla\mathcal{L}(\mathbf{X})\right) \right\|_2,\ \eta_g \in [a, b]$$

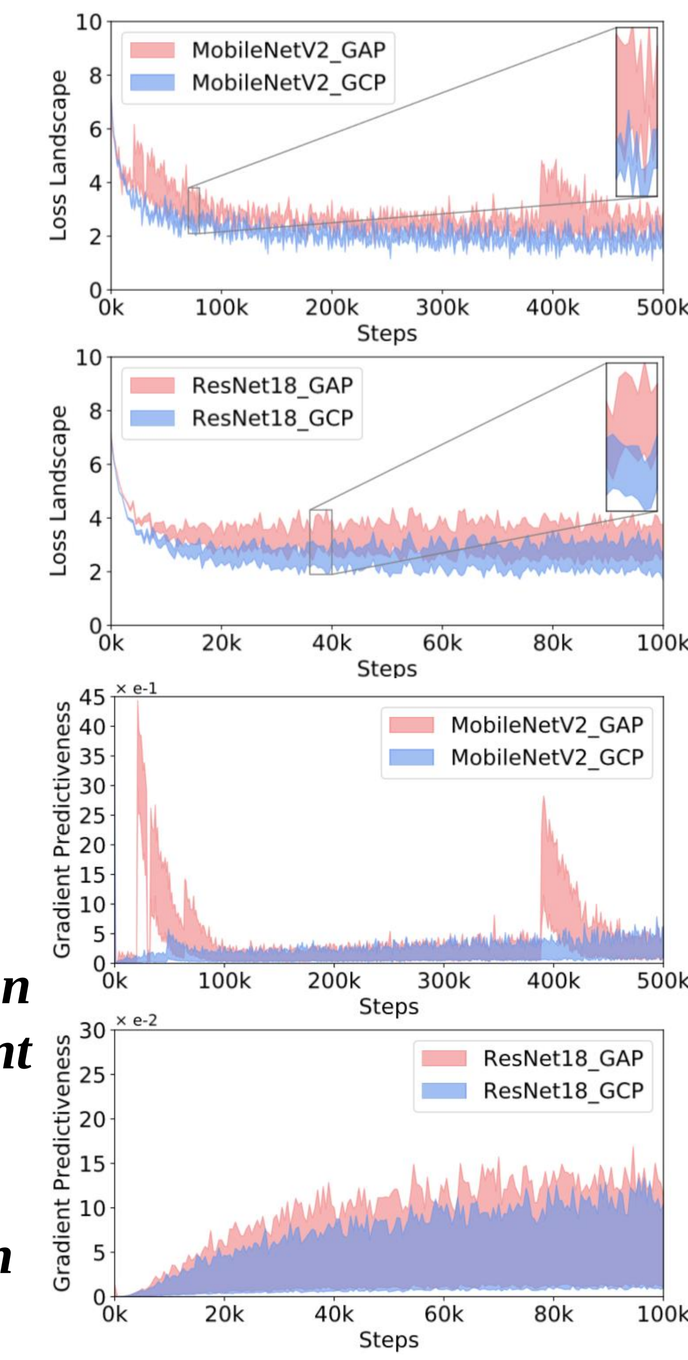
■ ***Results:***

- ***Networks with GCP have smaller variations of the optimization loss than GAP-based ones.***
- ***Gradients of networks with GCP are more stable than those of GAP-based ones.***

■ ***Conclusion:***

- ***GCP has the ability to smoothen optimization landscape of deep CNNs and improve gradient predictiveness.***

[1] Santurkar S , Tsipras D , Ilyas A , et al. How Does Batch Normalization Help Optimization?NeurIPS. 2018.



2. Connection to Second-order Optimization

Method	Gradient	Remark
GAP	$\eta\mathbf{C}^T \frac{\partial\mathcal{L}}{\partial\mathbf{Z}_{\text{GAP}}} \frac{\partial\mathbf{X}}{\partial\mathbf{W}_t}$	$\mathbf{C}$ is a constant matrix
GCP	$\approx \mathbf{F}^{-1} \frac{\partial\mathcal{L}}{\partial\mathbf{Z}_{\text{GCP}}} \frac{\partial\mathbf{X}}{\partial\mathbf{W}_t}$	$\mathbf{F}^{-1} = \eta 2\mathbf{J}\mathbf{X}\left(2\mathbf{K}^T \circ \mathbf{\Lambda}^{\frac{1}{2}} + \frac{1}{2}\mathbf{\Lambda}^{-\frac{1}{2}}\right)$
K-FAC [35]	$\mathbf{H}^{-1}\mathbf{C}^T \frac{\partial\mathcal{L}}{\partial\mathbf{Z}_{\text{GAP}}} \frac{\partial\mathbf{X}}{\partial\mathbf{W}_t}$	$\mathbf{H}^{-1} = \eta\left(\frac{\partial\mathcal{L}}{\partial\mathbf{X}}\right)^{-1} \otimes \widehat{\mathbf{X}}^{-1}$

Comparison of gradients involved in GAP, GCP and GAP with K-FAC

★ ***K-FAC is a second-order optimization method.***

■ ***The relationship between GCP and K-FAC:***

- ***Inverse of Hessian matrix: (GCP -output X and its eigenvalues) vs. (K-FAC-input and the gradient of output X).***
- ***The trimmed BP of GCP shares some similar philosophy with K-FAC.***

■ ***BP of GCP is a potential alternative of Hessian pre-conditioner.***

Merits Benefited from GCP

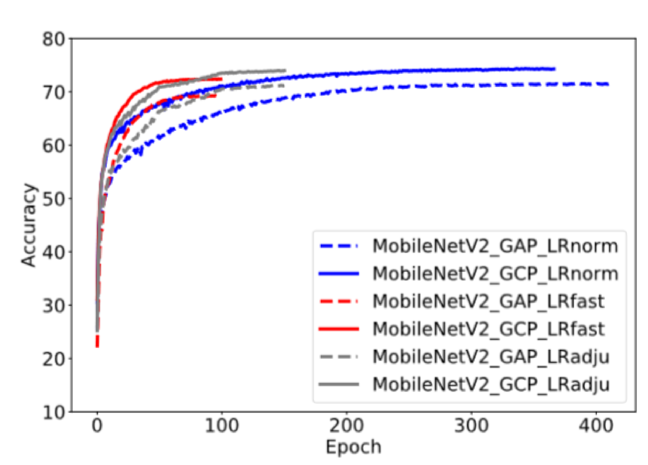
1. Acceleration of Network Convergence

Model	Method	lr	BS	Training Epochs	Matching Epoch	Top-1 Acc.	Top-5 Acc.
MobileNetV2	GAP	LR_{norm}	96	400	N/A	71.58	90.30
	GCP	LR_{adju}	192	150	$68_{(\downarrow 332)}$	$73.97_{(\uparrow 2.39)}$	$91.54_{(\uparrow 1.24)}$
ShuffleNetV2	GAP	LR_{norm}	1,024	240	N/A	67.96	87.84
	GCP	LR_{adju}	1,024	100	$78_{(\downarrow 162)}$	$71.17_{(\uparrow 3.21)}$	$89.74_{(\uparrow 1.90)}$
ResNet-18	GAP	LR_{norm}	256	100	N/A	70.47	89.62
	GCP	LR_{adju}	256	50	$32_{(\downarrow 68)}$	$74.86_{(\uparrow 4.39)}$	$91.81_{(\uparrow 2.19)}$
ResNet-34	GAP	LR_{norm}	256	100	N/A	74.19	91.61
	GCP	LR_{adju}	256	50	$38_{(\downarrow 62)}$	$76.81_{(\uparrow 2.62)}$	$93.09_{(\uparrow 1.48)}$
ResNet-50	GAP	LR_{norm}	256	100	N/A	76.17	92.93
	GCP	LR_{adju}	256	50	$40_{(\downarrow 60)}$	$78.03_{(\uparrow 1.86)}$	$93.95_{(\uparrow 1.02)}$
ResNet-101	GAP	LR_{norm}	256	100	N/A	77.67	93.89
	GCP	LR_{adju}	256	50	$41_{(\downarrow 59)}$	$79.18_{(\uparrow 1.51)}$	$94.51_{(\uparrow 0.62)}$

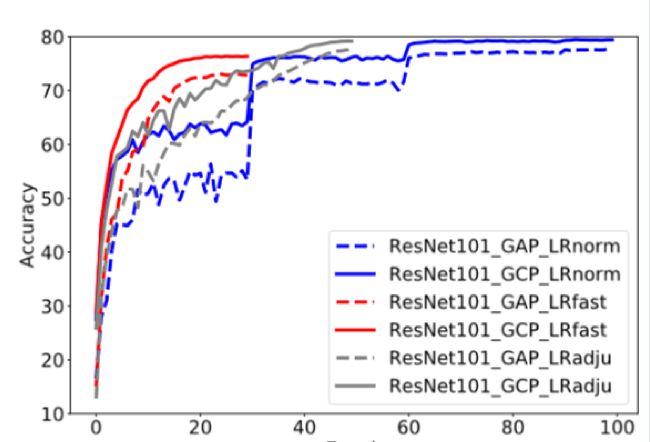
Comparison of model trained with GAP using LRnorm and with GCP using LRadju on ImageNet

■ ***GCP can significantly speed up convergence of deep CNNs with rapid decay of learning rates.***

- ***GCP achieves matching accuracies to GAP using only about $\frac{1}{3}$ training epochs.***
- ***GCP achieves 1.5%~4.4% accuracy improvement over GAP using less than $\frac{1}{2}$ training epochs.***



MobileNetV2



ResNet-101

2. Robustness to Distorted Examples

Method	IMAGENET-C		IMAGENET-P	
	mCE	Relative mCE	mFP	mT5D
MobileNetV2+GAP	87.1	114.9	79.8	96.5
MobileNetV2+GCP	$81.7_{(\downarrow 5.4)}$	$110.6_{(\downarrow 4.3)}$	$64.3_{(\downarrow 15.5)}$	$87.6_{(\downarrow 8.9)}$
ShuffleNetV2+GAP	92.7	126.7	94.7	108.2
ShuffleNetV2+GCP	$85.2_{(\downarrow 7.5)}$	$112.6_{(\downarrow 14.1)}$	$75.2_{(\downarrow 19.5)}$	$95.5_{(\downarrow 12.7)}$
ResNet-18+GAP	84.7	103.9	72.8	87.0
ResNet-18+GCP	$76.3_{(\downarrow 8.4)}$	$101.3_{(\downarrow 2.6)}$	$53.2_{(\downarrow 19.6)}$	$77.1_{(\downarrow 9.9)}$
ResNet-34+GAP	77.9	98.7	61.7	79.5
ResNet-34+GCP	$72.4_{(\downarrow 5.5)}$	$96.9_{(\downarrow 1.8)}$	$47.7_{(\downarrow 14.0)}$	$72.4_{(\downarrow 7.1)}$
ResNet-50+GAP	76.7	105.0	58.0	78.3
ResNet-50+GCP	$70.7_{(\downarrow 6.0)}$	$97.9_{(\downarrow 7.1)}$	$47.5_{(\downarrow 10.5)}$	$74.6_{(\downarrow 3.7)}$
ResNet-101+GAP	70.3	93.7	52.6	73.9
ResNet-101+GCP	$65.5_{(\downarrow 4.8)}$	$89.1_{(\downarrow 4.6)}$	$42.1_{(\downarrow 10.5)}$	$68.3_{(\downarrow 5.6)}$

Comparison of GAP and GCP on IMAGENET-C and IMAGENET-P

■ ***GCP can greatly improve the robustness of deep CNNs to common image corruptions and perturbations.***

- ***5~8.5 and 2~14 improvement on ImageNet-C.***
- ***10~20 and 4~13 improvement on ImageNet-P.***

3. Generalization Ability to Other Tasks

Backbone Model	Method	Detectors	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet-50	GAP	Faster R-CNN	36.4	58.2	39.2	21.8	40.0	46.2
	GCP_D		$36.6_{(\uparrow 0.2)}$	$58.4_{(\uparrow 0.2)}$	$39.5_{(\uparrow 0.3)}$	$21.3_{(\downarrow 0.5)}$	$40.8_{(\uparrow 0.8)}$	$47.0_{(\uparrow 0.8)}$
	GCP_M		$37.1_{(\uparrow 0.7)}$	$59.1_{(\uparrow 0.9)}$	$39.9_{(\uparrow 0.7)}$	$22.0_{(\uparrow 0.2)}$	$40.9_{(\uparrow 0.9)}$	$47.6_{(\uparrow 1.4)}$
ResNet-101	GAP		38.7	60.6	41.9	22.7	43.2	50.4
	GCP_D		$39.5_{(\uparrow 0.8)}$	$60.7_{(\uparrow 0.1)}$	$43.1_{(\uparrow 1.2)}$	$22.9_{(\uparrow 0.2)}$	$44.1_{(\uparrow 0.9)}$	$51.4_{(\uparrow 1.0)}$
	GCP_M		$39.6_{(\uparrow 0.9)}$	$61.2_{(\uparrow 0.6)}$	$43.1_{(\uparrow 1.2)}$	$23.3_{(\uparrow 0.6)}$	$43.9_{(\uparrow 0.7)}$	$51.3_{(\uparrow 0.9)}$
ResNet-50	GAP	Mask R-CNN	37.2	58.9	40.3	22.2	40.7	48.0
	GCP_D		$37.3_{(\uparrow 0.1)}$	$58.8_{(\downarrow 0.1)}$	$40.4_{(\uparrow 0.1)}$	$22.0_{(\downarrow 0.2)}$	$41.1_{(\uparrow 0.4)}$	$48.2_{(\uparrow 0.2)}$
	GCP_M		$37.9_{(\uparrow 0.7)}$	$59.4_{(\uparrow 0.5)}$	$41.3_{(\uparrow 1.0)}$	$22.4_{(\uparrow 0.2)}$	$41.5_{(\uparrow 0.8)}$	$49.0_{(\uparrow 1.0)}$
ResNet-101	GAP		39.4	60.9	43.3	23.0	43.7	51.4
	GCP_D		$40.3_{(\uparrow 0.9)}$	$61.5_{(\uparrow 0.6)}$	$44.0_{(\uparrow 0.7)}$	$24.1_{(\uparrow 1.1)}$	$44.7_{(\uparrow 1.0)}$	$52.5_{(\uparrow 1.1)}$
	GCP_M		$40.7_{(\uparrow 1.3)}$	$62.0_{(\uparrow 1.1)}$	$44.6_{(\uparrow 1.3)}$	$23.9_{(\uparrow 0.9)}$	$45.2_{(\uparrow 1.5)}$	$52.9_{(\uparrow 1.5)}$

Object detection of various deep CNN models using Faster R-CNN and Mask R-CNN on COCO val2017

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
R-50+GAP	34.1	55.5	36.2	16.1	36.7	50.0
R-50+GCP_D	34.2	55.3	36.4	15.8	37.1	50.1
R-50+GCP_M	34.7	56.3	36.8	16.4	37.5	50.6
R-101+GAP	35.9	57.7	38.4	16.8	39.9	53.5
R-101+GCP_D	36.5	58.2	39.1	17.6	39.9	53.5
R-101+GCP_M	36.7	58.7	39.1	17.6	39.9	53.7

Instance segmentation of various deep CNN models using Mask R-CNN on COCO val2017

■ ***GCP has good generalization ability to other tasks.***

- ***GCP improves ~0.9% over GAP on object detection.***
- ***GCP improves ~0.8% over GAP on instance segmentation.***