Synergistic Self-supervised and Quantization Learning

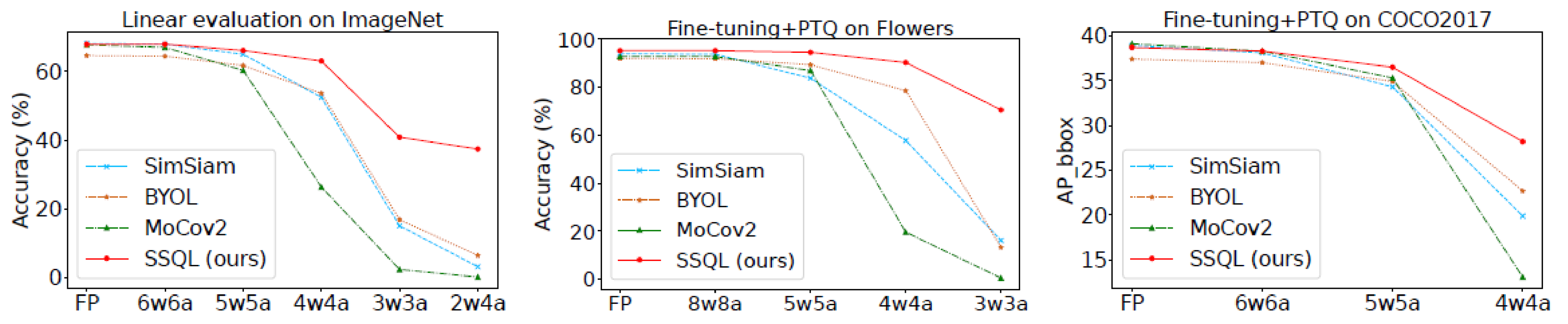
Yun-Hao Cao[1], Peiqin Sun[2*], Yechang Huang[2], Jianxin Wu[1], Shuchang Zhou[2]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]MEGVII Technology

TEL AVIV 2022

1. Introduction & Motivation

In this paper, we propose a method called synergistic self-supervised and quantization learning (SSQL) to pretrain quantization-friendly self-supervised models facilitating downstream deployment.

- With the fast development of self-supervised learning (SSL), an increasing proportion of the models that need to be deployed in downstream tasks are finetuned from SSL pretrained models.
- To facilitate deployment, quantization is one of the most effective methods and is directly supported by most current hardware.
- Current state-of-the-art SSL methods all incur severe drop in accuracy when bit-width goes below 5.



(a) ImageNet (b) Flowers (c) COCO2017 detection

Key idea

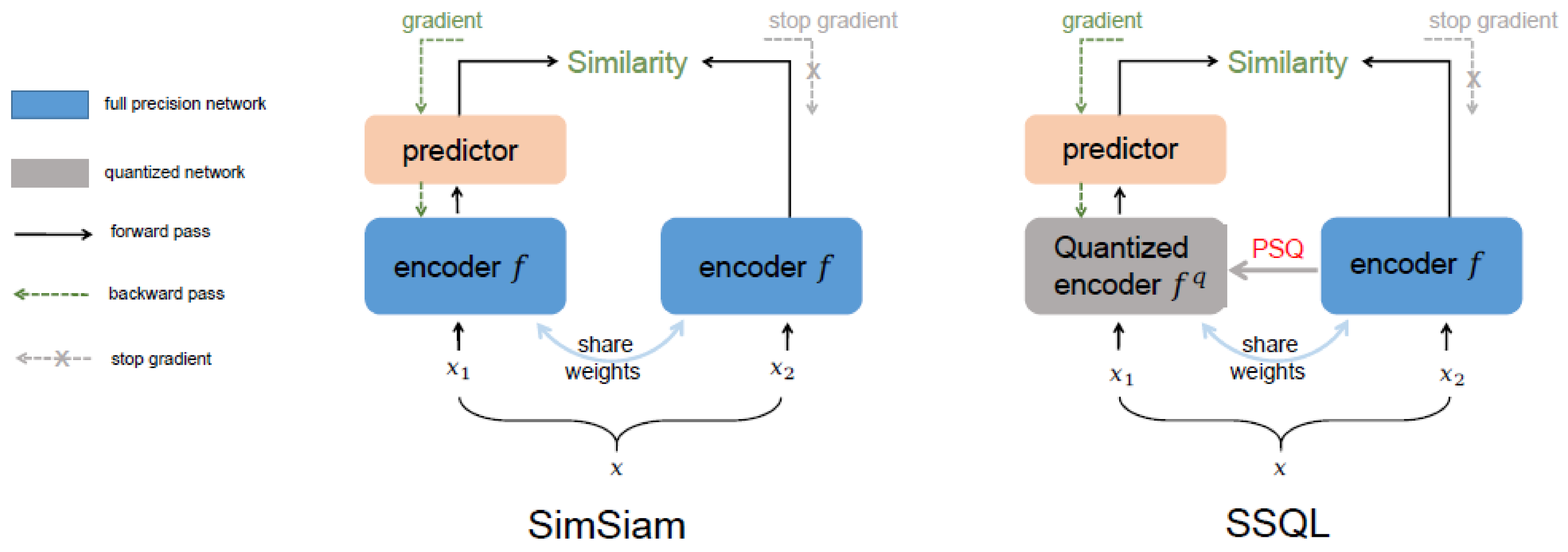
- Can we learn a quantization-friendly representation such that the pretrained model can be quantized more easily to facilitate deployment when transferring to various downstream tasks?

Key Property

- ➢ Train only once
- ➢ One copy of weights
- ➢ Bit-width flexibility
- ➢ Improve the accuracy of full precision models in most cases

2. Framework of proposed SSQL

Our SSQL can be combined with various SSL methods and here we use SimSiam as the baseline.



$$L_{SSQL} = D(\mathbf{p}_1^q, SG(\mathbf{z}_2)) + D(\mathbf{p}_2^q, SG(\mathbf{z}_1))$$

➢ How to quantize the network during training?

We propose PSQ, in which we calculate S and Z in each step, to adapt to the changing weights.

➢ How to update the weights?

The quantized network f_q and FP network f share weights and we directly backprop the quantized network f_q using STE.

➢ How to achieve bit-width flexibility?

We randomly select values from a set of candidate bit-widths in each step for the assignment of q.

The synergy between SSL and quantization

$$\begin{aligned}
&\mathbb{E}_{x,\mathcal{T},q}\left[\|\mathcal{F}_\theta^q(\mathcal{T}(x)) - \mathcal{F}_{\theta^t}(\mathcal{T}'(x))\|_2^2\right] \\
=&\mathbb{E}_{x,\mathcal{T},q}\left[\|\mathcal{F}_\theta^q(\mathcal{T}(x)) - \mathcal{F}_\theta(\mathcal{T}(x)) + \mathcal{F}_\theta(\mathcal{T}(x)) - \mathcal{F}_{\theta^t}(\mathcal{T}'(x))\|_2^2\right] \\
=&\underbrace{\mathbb{E}_{x,\mathcal{T},q}\left[\|\mathcal{F}_\theta^q(\mathcal{T}(x)) - \mathcal{F}_\theta(\mathcal{T}(x))\|_2^2\right]}_{\text{Q term (quantization term)}} + \underbrace{\mathbb{E}_{x,\mathcal{T},q}\left[\|\mathcal{F}_\theta(\mathcal{T}(x)) - \mathcal{F}_{\theta^t}(\mathcal{T}'(x))\|_2^2\right]}_{\text{CL term (contrastive learning term)}} + \underbrace{2\mathbb{E}_{x,\mathcal{T},q}\left[\left(\mathcal{F}_\theta^q(\mathcal{T}(x)) - \mathcal{F}_\theta(\mathcal{T}(x))\right)^T\left(\mathcal{F}_\theta(\mathcal{T}(x)) - \mathcal{F}_{\theta^t}(\mathcal{T}'(x))\right)\right]}_{\text{cross term}}
\end{aligned}$$

3. Experiments

CIFAR-10 results

Backbone	Method	Linear evaluation accuracy (%) FP	8w8a	6w6a	5w5a	4w4a	3w3a	2w8a	2w4a
ResNet-18	SimSiam [6]	**90.7**	90.7	**90.6**	90.3	88.9	66.0	70.1	63.8
	BYOL [13]	89.3	89.3	89.4	89.3	88.0	75.1	71.9	63.3
	SimSiam-PACT [7]	-	89.2	89.2	89.3	89.2	88.2	89.3	88.3
	SSQL (ours)	**90.7**	**90.8**	**90.6**	**90.6**	**90.1**	**85.6**	**88.0**	**86.5**
	SimCLR [4]	**89.4**	**89.3**	**89.2**	**88.8**	87.1	73.9	65.6	55.6
	MoCov2 [5]	88.9	88.8	88.4	88.2	86.8	72.2	66.4	50.7
	SSQL-NCE (ours)	89.0	89.0	89.0	**88.8**	**87.9**	**82.9**	**87.1**	**84.9**
ResNet-50	SimSiam [6]	90.9	90.9	91.0	90.6	89.5	74.1	55.1	57.1
	BYOL [13]	90.3	90.3	90.0	89.7	87.5	58.5	82.4	67.8
	SSQL (ours)	**91.1**	**91.1**	**91.1**	**91.1**	**90.0**	**77.4**	**89.5**	**87.2**
	SimCLR [4]	91.5	91.4	91.3	90.5	88.1	59.6	63.5	42.4
	MoCov2 [5]	90.2	90.2	90.2	89.4	87.9	72.1	68.8	49.5
	SSQL-NCE (ours)	**92.1**	**92.1**	**92.0**	**91.9**	**89.8**	**74.0**	**88.6**	**84.9**

- SSQL achieves better performance than counterparts in most bit-widths.

ImageNet results

Backbone	Method	Linear evaluation accuracy (%) FP	8w8a	5w5a	4w4a	3w3a	2w4a
ResNet-18	SimSiam [6]	55.0	54.7	53.9	36.7	6.3	1.5
	BYOL [13]	54.1	54.0	51.9	42.4	13.6	3.6
	SimSiam-PACT [7]	-	52.8	52.8	52.3	51.0	51.6
	SSQL (ours)	**57.6**	**57.6**	**56.7**	**52.8**	**41.0**	**43.1**
ResNet-50	SimSiam [6]	**68.1**	**67.9**	65.0	52.4	15.0	3.1
	BYOL [13]	64.6	64.4	61.7	53.6	16.8	6.4
	MoCov2† [5]	67.7	67.0	60.3	26.3	2.3	0.1
	SSQL (ours)	67.9	**67.9**	**66.1**	**63.0**	**40.8**	**37.4**

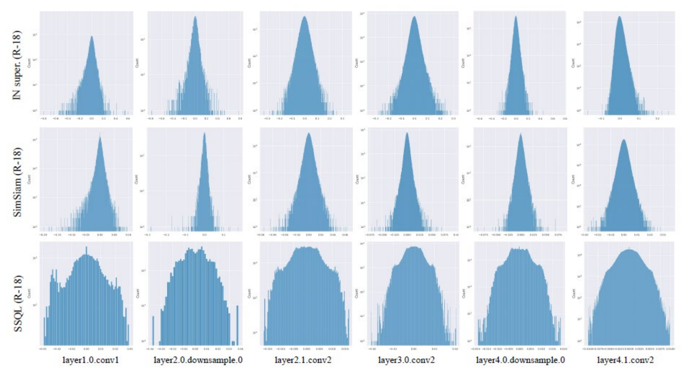
- Comparable FP performance, together with much better performance at lower bit-widths.

Object detection&segmentation on COCO

Method	FP AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}	6w6a AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
IN supervised	38.2	56.0	42.0	34.8	56.0	37.2	37.6	58.3	41.4	34.3	55.2	36.8
SimSiam	**38.9**	**59.8**	**42.3**	**35.2**	**56.7**	**37.7**	38.1	58.7	41.5	34.5	55.7	36.8
BYOL	37.4	57.9	40.6	34.1	54.9	36.4	37.0	57.4	40.2	33.7	54.3	36.0
SSQL (ours)	38.7	59.2	**42.3**	**35.2**	56.2	**37.7**	**38.3**	**58.8**	**41.7**	**34.8**	**55.8**	**37.3**
	5w5a						4w4a					
IN supervised	35.2	55.5	38.4	31.9	52.3	34.0	23.4	38.6	24.6	21.4	36.3	22.1
SimSiam	34.3	54.0	36.7	30.9	50.6	32.6	19.9	33.6	20.6	18.1	31.3	18.3
BYOL	34.9	54.4	37.7	31.8	51.4	33.8	22.7	37.4	24.0	20.9	35.2	21.7
SSQL (ours)	**36.5**	**56.9**	**39.4**	**33.3**	**53.6**	**35.5**	**28.2**	**43.1**	**27.5**	**26.0**	**43.1**	**27.5**

- SSQL surpasses the other methods at all bit-widths by a large margin even after fine-tuning.

Weight distribution



- Smaller dynamic ranges
- More uniformly distributed
- Fewer outliers

4. Contributions & Conclusions

✓ We design an effective method called SSQL, which not only greatly improves the performance when quantized to low bit-widths, but also boosts the performance of full precision models in most cases.

✓ With SSQL, models only need to be trained once and can then be customized for a variety of downstream tasks at different bit-widths.

✓ We provide theoretical analysis about the synergy between SSL and quantization in SSQL.

✓ Exhaustive experimental results further show that our SSQL achieves better performance on various benchmarks at all bit-widths.