TOSHIBA BMVC 2023

SSCQ: Hierarchical Quantization Consistency for Fully Unsupervised Image Retrieval

Guile Wu Chao Zhang Stephan Liwicki
https://github.com/cazhang/sscq

Motivations

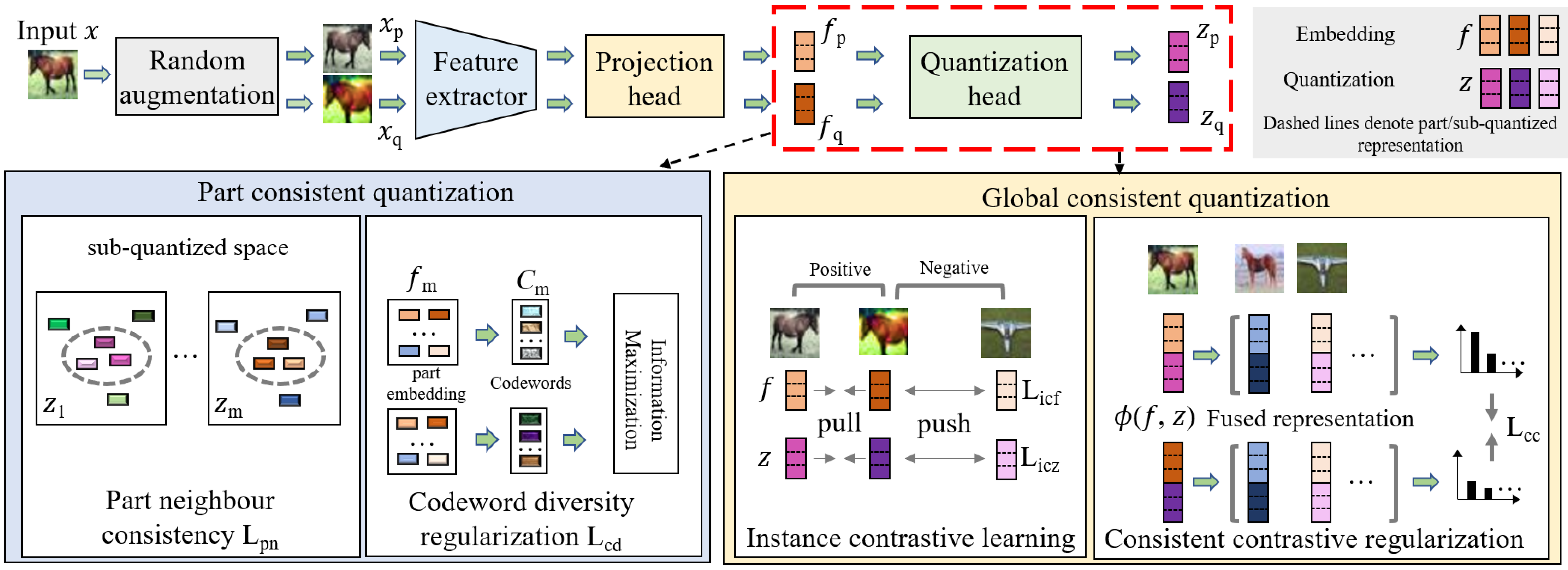
- Unsupervised image retrieval works *without* data annotations
- Existing methods using self-supervised learning
- We tackle false negative issue of contrastive loss

Proposed method

- Exploit sub-quantized representations for self-supervised learning
- Leverage consistency to regularize the instance contrastive learning
- With a unified objective, our approach exploits richer self-supervision cues

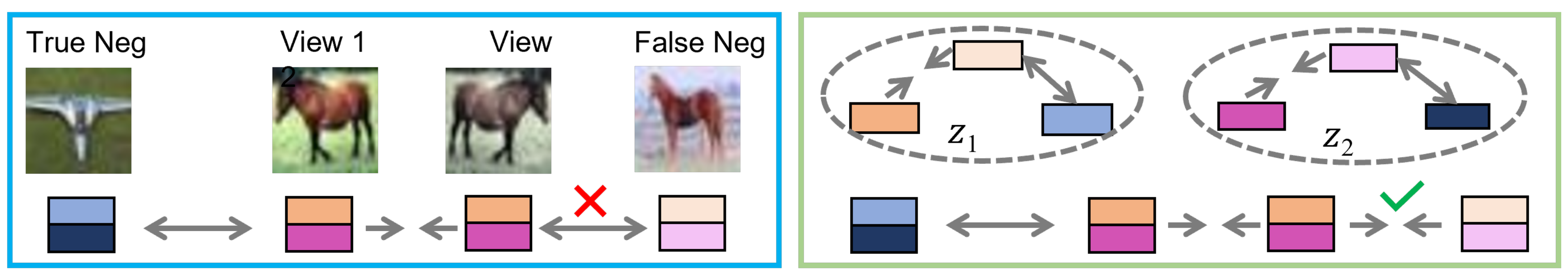
Contributions

- Propose a hierarchical consistent quantization approach for deep fully unsupervised image retrieval
- Global: improve retrieval performance by exploiting contrastive consistency
- Part: employ neighbor semantic consistency learning in a self-supervised way



An overview of the proposed Self-Supervised Consistent Quantization (SSCQ) approach to deep fully unsupervised image retrieval. Part consistent quantization discovers part neighbor affinity as self-supervision, while global consistent quantization learns instance affinity as self-supervision, which together are formulated into a unified learning objective for model optimization.

Motivational example



(a) Instance contrastive loss makes mistakes for false negative (b) Part semantic loss reduces false negative error

(a) Given two views of the query instance of a *horse*, we illustrate the benefit of using part semantic loss with a true negative (*plane*) and a false negative (*another horse*). In (a), the instance contrastive loss with false negatives leads to sub-optimal feature representation. In (b), part embeddings of the anchor instance could be pulled closer to those from the *other horse*, thereby fixing the error caused by false negative in (a).

Proposed loss terms

- Instance contrastive learning loss:

$$\mathcal{L}_{icz} = -\log \frac{\exp(s(z, z^{+})/\tau_{ic})}{\sum_{j=1}^{2N_b} \mathbf{1}_{[z_j \neq z]} \exp(s(z, z_j)/\tau_{ic})}, \tag{1}$$

- Part Semantic Consistent Quantization:

$$\mathcal{L}_{pn} = -\frac{1}{M}\sum_{m=1}^{M} \log \frac{\sum_{n=1}^{N_k} \exp(s(z_m, z^{-}_{m,n})/\tau_{pn})}{\sum_{j=1}^{2Nb-2} \exp(s(z_m, z^{-}_{m,j})/\tau_{pn})}, \tag{2}$$

- Global Affinity Consistent Quantization:

$$Q(i) = \frac{\exp(s(\Phi(f,z), \Phi(f^{-}, z^{-})_i)/\tau_{cc})}{\sum_{j=1}^{2N_b-2} \exp(s(\Phi(f,z), \Phi(f^{-}, z^{-})_j)/\tau_{cc})},$$
$$P(i) = \frac{\exp(s(\Phi(f^{+},z^{+}), \Phi(f^{-}, z^{-})_i)/\tau_{cc})}{\sum_{j=1}^{2N_b-2} \exp(s(\Phi(f^{+},z^{+}), \Phi(f^{-}, z^{-})_j)/\tau_{cc})}, \tag{3}$$

Thus, contrastive consistency loss $\mathcal{L}_{cc}$ is defined using the symmetric KL Divergence D_{KL}, as:

$$\mathcal{L}_{cc} = \frac{1}{2}(D_{KL}(P\|Q) + D_{KL}(Q\|P)). \tag{4}$$

- Summary:

$$\mathcal{L} = \mathcal{L}_{icz} + \mathcal{L}_{icf} + \lambda_{pn}\mathcal{L}_{pn} + \lambda_{cd}\mathcal{L}_{cd} + \lambda_{cc}\mathcal{L}_{cc}, \tag{5}$$

Comparison with the State of the Art

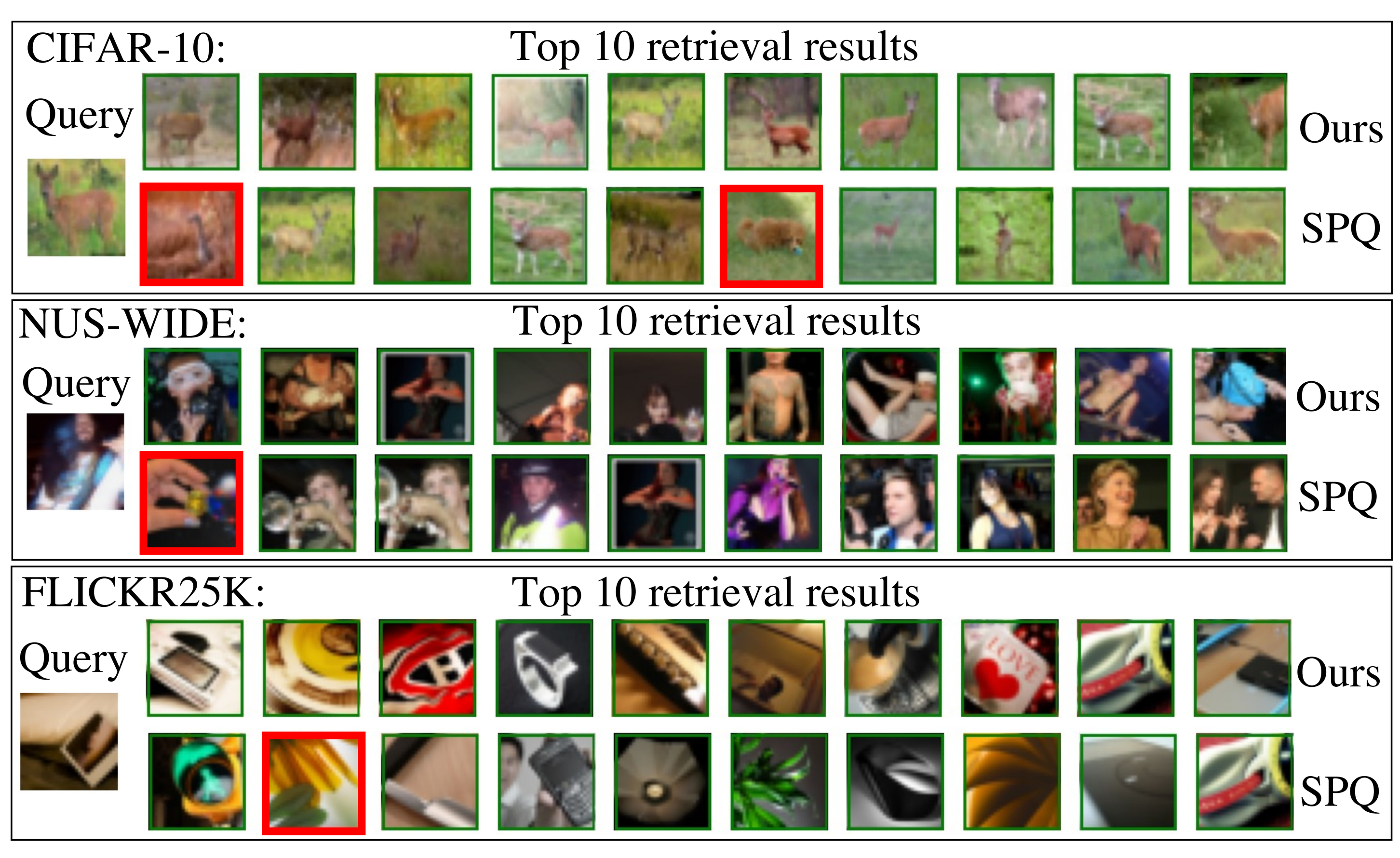
Dataset	Method	16 bits	32 bits	64 bits
CIFAR-10	SGH [Dai 2017]	43.5	43.7	43.3
	HashGAN [Dizaji 2018]	44.7	46.3	48.1
	BinGAN [Zieba 2018]	47.6	51.2	52.0
	SPQ [Jang 2021]	76.8	79.3	81.2
	SSCQ (ours)	**78.3**	**81.3**	**82.9**
NUS-WIDE	SGH [Dai 2017]	59.3	59.0	60.7
	HashGAN [Dizaji 2018]	68.4	70.6	71.7
	BinGAN [Zieba 2018]	65.4	70.9	71.3
	SPQ† [Jang 2021]	75.7	79.4	80.2
	SSCQ (ours)	**78.7**	**79.9**	**80.8**
FLICKR25K	SPQ [Jang 2021]	71.8	74.0	74.5
	SSCQ (ours)	**73.8**	**75.9**	**76.7**

Comparison with SOTA deep fully unsupervised methods on CIFAR-10, NUS-WIDE and FLICKR25K in terms of mAP (%).

Coupling part loss with global losses

Global Loss	$\mathcal{L}_{pn}$	mAP(%)↑	SimPos↑	SimNeg↓	Margin↑
$\mathcal{L}_{icz}$	-	74.48	0.68	0.09	0.59
	✓	77.25	0.72	0.10	0.62
$\mathcal{L}_{icf}$	-	10.59	0.29	-0.01	0.30
	✓	76.11	0.29	-0.03	0.32
$\mathcal{L}_{icz} + \mathcal{L}_{icf}$	-	76.28	0.30	-0.03	0.33
	✓	78.64	0.30	-0.03	0.33
SPQ[Jang 2021]	-	74.73	0.32	-0.03	0.35
	✓	74.96	0.32	-0.04	0.36

Qualitative visualizations



Retrieval results of our approach and SPQ [Jang 2021] on CIFAR-10, NUS-WIDE and FLICKR25K (32 bits). False retrieval results are denoted in red bounding boxes.