Exploit Clues from Views:
Self-Supervised and Regularized Learning for Multiview Object Recognition

Chih-Hui Ho, Bo Liu, Tz-Ying Wu, Nuno Vasconcelos
University of California, San Diego

Statistical Visual Computing Lab
UC San Diego

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Introduction

- Multiview recognition has been well studied in the literature and achieves decent performance in object recognition and retrieval task.
- However, most previous works rely on supervised learning and some impractical underlying assumptions
  - The availability of all views in training and inference time.
  - View labels, consistent view angles or the same number of object views are required.
- These limitations prevent many applications of interest. For example, the setting of Fig. 1, where a household robot of limited memory is tasked with picking scattered objects and returning them to their locations.
- We refer this problem as lightweight unsupervised multiview object recognition, which requires the algorithm to be fast convergence, memory efficient and object view invariant.
- To overcome this issue, we propose **multiview stochastic prototype embedding (MVSPE)** by encouraging views of an object naturally clustering around its object invariant.
- MVSPE is further regularized by minimizing the distribution shift when changing the set of view prototypes. The resulting embedding is referred as **view invariant stochastic prototype embedding (VISPE).**
- Experiments show that MVSPE and VISPE achieve good performance for retrieval and k nearest neighbor classification on (1) ModelNet [9] (2) ShapeNet [10] (3) sampled ModelNet with incomplete views.

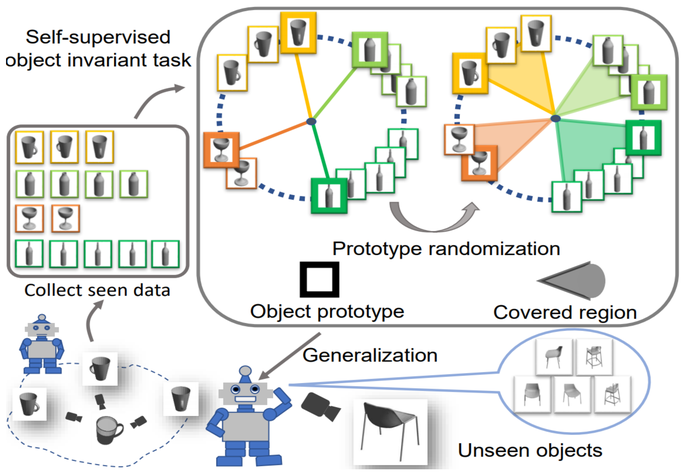

Figure1. Lightweight unsupervised multiview object recognition. A household robot collects multiple object views by moving around, aggregating a multiview object database without view labels. A self-supervised learning algorithm is applied to this database to create an embedding that maps images from same object into an object invariant. At inference time, this embedding generalizes to new views, objects, and object classes.

Code available at
https://github.com/chihhuiho/VISPE

Proposed method

- Consider a set of objects $\mathrm{O} = \{o_i\}_{i=1}^{N}$ and each object consists of images from V_i unspecified viewpoints $o_i = \{x_i^j\}_{j=1}^{V_i}$.
- For self-supervised learning, each object o_i is treated as a different class, establishing a labelled image dataset $D = \{(x_i^j, y_i^j) | y_i^j = i, \forall j \in V_i\}$.
- A naïve baseline is to implement a instance classifier with a softmax layer $P_{Y|X}(i|x) = \frac{\exp(W_i^T f_\theta(x))}{\sum_{k=1}^{N} \exp(W_k^T f_\theta(x))}$, where w_i is the parameter vector of instance i. This is illustrated in Figure 2(a).
- However, the softmax classifier is most successful for closed-set classification, where train and test object classes are the same. In general, the learned embedding f_θ does not have a good metric structure beyond these classes.
- To forces the embedding to have a good metric structure over larger regions of the feature space, we consider randomization strategies that leverage the view richness of multiview datasets to achieve better generalization to unseen classes during training.
- This is implemented by replacing the classifier weight w_i with the normalized embedding of a randomly chosen view of object instance o_i, as shown in Figure 2(b).
- The random view of an object is selected by the view sampler v_i following the random schedule algorithm. See algorithm 1 and Figure 2(c-e) for more details.
- With the set s of selected views, multiview stochastic prototype embedding (**MVSPE**) is defined as

$$P^s_{Y|X}(i|x) = \frac{exp(f_\theta(x_i^{v_i})^T f_\theta(x)/\tau)}{\sum_{k=1}^{N} exp(f_\theta(x_k^{v_k})^T f_\theta(x)/\tau)},$$

  where τ is the temperature and the loss L_s is defined as $L_s = -\log(P^s_{Y|X}(i|x))$.
- To encourage the classifier parameters remaining stable under different sets s_1 and s_2 of selected views, KL divergence is proposed to regularized the probability shift when using s_1 and s_2 as

$$L_{KL} = \sum_{k=1}^{m} P^{s_1}_{Y|X}(i|x) log \frac{P^{s_1}_{Y|X}(i|x)}{P^{s_2}_{Y|X}(i|x)}.$$

- The view invariant stochastic prototype embedding (**VISPE**) is trained with loss function $L = L_{s_1} + L_{s_2} + \alpha L_{KL}$ with $\tau = 0.05$ and $\alpha = 5$.

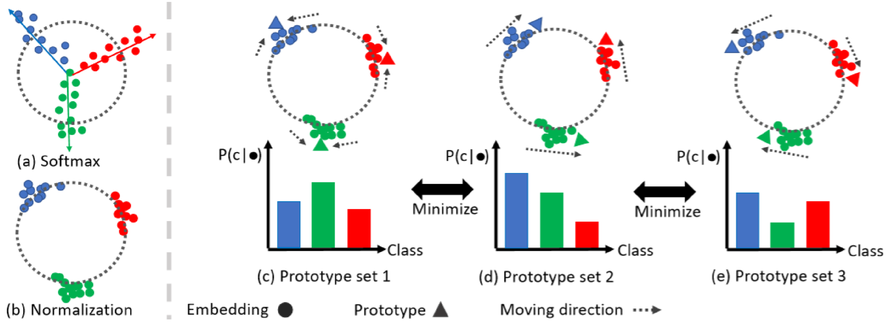

Figure 2: Regularization of a self-supervised embedding by prototype randomization. Each color represents a single object and views of the same object are marked with same color. (a) softmax embedding, unnormalized features. Solid arrows represent the weight vectors w_i learned per instance i. (b) Normalized embedding. (c-e) Randomization: 3 different sets of prototypes are used for training. Dashed lines show how view embeddings are encouraged to move towards the object prototype. Bar plots illustrate the posterior class probabilities of a given image changse when the prototypes are switched. The proposed VISPE minimizes these variations.

Algorithm 1 Randomization schedule

Input Threshold t
Use the view samplers $\nu_i, \forall i$ to select a set of random prototypes $\mathcal{W} = \{f_\theta(x_1^{\nu_1}), \ldots, f_\theta(x_N^{\nu_N})\}$ to use in (3).
while Not convergence **do**
  Minimize the risk of (2)
  for all $i \in N$ **do**
    $u \sim Unif(0, 1)$
    if $u < t$ **then**
      Use ν_i to resample a new prototype $f_\theta(x_i^{\nu_i})$
      $w_i \leftarrow f_\theta(x_i^{\nu_i})$
    end if
  end for
end while

Experiments

Table 1: KNN classification results for various baselines, solving different surrogate tasks. RSPE outperforms all self-supervised learning methods, VGG16 [1] pretrained model and instance classifiers.

Datasets	Surrogate	ModelNet		ShapeNet		ModelNet-S	
Methods / Classes	Task	seen	unseen	seen	unseen	seen	unseen
Chance	N/A	3.3	10.0	3.3	4.0	3.3	10.0
Pretrained [1]	N/A	62.7	52.7	63.9	58.1	58.2	55.2
Autoencoder [2]	Context	31.8	37.2	29.8	26.3	34.7	38.8
Egomotion [3]	Motion	32.4	34.7	72.6	47.1	33.0	35.2
Puzzle [4]	Sequence	34.4	41.5	67.8	48.6	34.8	42.4
UEL [5]	Data Aug.	47.9	46.5	68.7	53.4	46.4	48.2
ShapeCode [6]	View	39.4	46.5	67.1	42.3	38.8	47.2
MVCNN [7]	N/A	39.6	48.1	30.3	32.4	36.7	44.8
Triplet [8]	N/A	70.1	62.4	81.2	61.2	64.7	62.1
Instance classifier	N/A	57.7	58.9	69.3	60.4	52.3	54.6
PE	Object	69.7	61.7	81.6	63.8	62.1	60.4
MVSPE	Object	70.3	63.2	82.4	64.6	64.6	62.1
VISPE	Object	**71.2**	**64.4**	**82.9**	**65.5**	**66.2**	**64.3**

Table 2: Left: Recall @ k and NMI on Modelnet unseen classes. Right: low shot accuracy for k labeled images.

	Retrieval and Clustering Recall					Low shot k Images		
Methods	@1	@2	@4	@8	NMI	1	3	5
Pretrained [1]	94.5	96.6	98.2	**99.3**	46.7	34.3	46.8	51.2
Autoencoder [2]	81.7	86.8	92.3	95.2	25.4	25.0	30.0	28.0
Egomotion [3]	73.4	80.7	88.0	92.9	7.5	15.1	18.1	19.9
Puzzle [4]	77.8	84.1	89.8	94.0	21.9	21.1	26.8	29.3
UEL [5]	77.8	85.4	91.6	95.7	24.6	23.8	30.9	34.2
ShapeCode [6]	83.4	88.5	93.4	96.2	27.4	28.8	36.1	39.5
MVCNN [7]	80.3	86.7	91.7	95.0	19.3	21.6	27.0	29.5
Triplet [8]	90.8	94.7	97.4	98.8	48.2	41.4	50.3	54.5
Instance classifier	89.1	92.5	95.6	97.4	37.1	28.3	42.2	48.4
PE	91.2	95.0	97.2	98.5	48.2	40.2	49.7	52.9
MVSPE	92.4	95.4	97.7	98.9	48.4	41.5	50.8	54.2
VISPE	**95.5**	**97.7**	**98.6**	99.2	**51.1**	**43.1**	**52.5**	**55.9**

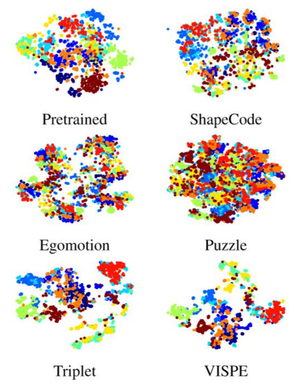

Figure 3: TSNE visualization of unseen class embeddings. Each color represents a class. RSPE produces more structured embedding.

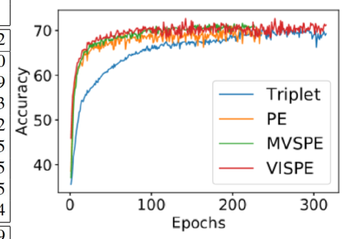

Figure 4: Convergence rate of proposed methods and triplet loss.

Conclusion

- The current impractical assumptions of supervised multiview recognition are discussed.
- To relax these assumptions, multiview stochastic prototype embedding (MVSPE) is proposed for learning object invariant representation in self-supervised manner.
- View invariant stochastic prototype embedding (VISPE) is further proposed to regularize the learning of the embedding and outperforms other baselines with faster convergence.

Acknowledgement

This work was partially funded by NSF awards IIS-1637941, IIS-1924937, and NVIDIA GPU donations.

References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014
[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT Press, 2016.
[3] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 37–45, Dec 2015.
[4] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. CoRR, abs/1603.09246, 2016
[5] Mang Ye, Xu Zhang, Pong C. Yuen, and Shih-Fu Chang. Unsupervised embedding learning via invariant and spreading instance feature. CoRR, abs/1904.03436, 2019.
[6] Dinesh Jayaraman, Ruohan Gao, and Kristen Grauman. Unsupervised learning through one-shot image-based shape reconstruction. CoRR, abs/1709.00505, 2017.
[7] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. CoRR, abs/1505.00880, 2015.
[8] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. CoRR, abs/1503.03832, 2015.
[9] Zhirong Wu, S. Song, A. Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1912–1920, June 2015.
[10] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. CoRR, abs/1512.03012, 2015.