Visual Similarity from Optimizing Feature and Memory On A Hypersphere

Xinlei Pan, Rudrasis Chakraborty, Stella X. Yu

UC Berkeley / ICSI

Objectives

We propose an unsupervised metric learning method that develops apparent visual similarity from images alone. Our method maps high-dimensional visual data onto a low-dimensional hyper-sphere and consolidate such feature representations into a visual memory representation. Optimizing the feature mapping and visual memory on a hypersphere achieves maximal discrimination among instances. We show through extensive experiments that our algorithm achieves better classification accuracy, convergence rate, and feature transferability, and can be useful for initializing policy networks for vision-based reinforcement learning tasks to improve sample efficiency.

Introduction

Supervised learning such as image classification implicitly finds visual similarity between images. Based on this inspiration, we explore an unsupervised learning approach that learns to perform instance-discriminative learning by maximizing feature distances of different images. We compare our method with Wu *et al.* [1] in several image classification tasks and visual reinforcement learning tasks.

Figure: Visualization of learned low dimensional feature representations on CIFAR10 dataset. Left: ours; Right: Baseline. The visualization is done using t-SNE dimension reduction. Different colors indicate different image classes.

Methods

Our model has two sets of parameters:

- the parametric feature mapping function $\mathbf{f}(x;\theta)$ that takes an image to a point on a unit hypersphere (denoted by $\mathbf{S}^m$);
- the non-parametric feature memory bank V that stores the consolidated representation for all the training instances. Let $\{x_i\}_{i=1}^n$ be the set of images.

We use $\mathbf{f}_i = \mathbf{f}(x_i;\theta)$ to denote the feature corresponds to image x_i. $\|\mathbf{f}_i\| = 1 \forall x_i$.
With image x mapped to feature $\mathbf{f}(x;\theta)$, the probability of x as an observation of instance class $\mathbf{v}_i$ in memory depends on the distance between x and $\mathbf{v}_i$ and is given by:

$$P(x;\theta,V) = \frac{\exp(-d^2(\mathbf{f}(x;\theta),\mathbf{v}_i)/T)}{\sum_{j=1}^n \exp(-d^2(\mathbf{f}(x;\theta),\mathbf{v}_j)/T)}, \quad (1)$$

where, T is the tunable hyperparameter. Assume that $x_1,\cdots,x_n$ are i.i.d. samples. The joint probability of drawing samples $x_1,\cdots,x_n$ is given by:

$$\begin{aligned} P(x_1,\ldots,x_n;\theta,V) &= \prod_{i=1}^n P(x_i;\theta,V) \\ &= \prod_{i=1}^n \frac{\exp(-d^2(\mathbf{f}_i,\mathbf{v}_i)/T)}{\sum_{j=1}^n \exp(-d^2(\mathbf{f}_i,\mathbf{v}_j)/T)}. \end{aligned} \quad (2)$$

The optimal feature mapping θ^* and visual memory $V*$ should be obtained by maximizing the above likelihood or equivalently minimizing the negative log likelihood $\ell(\theta,V)$ as given below:

$$\begin{aligned} (\theta^*,V^*) &= \arg\max_{\theta,V} P(x_1,\ldots,x_n;\theta,V) \\ &= \arg\min_{\theta,V} \ell(\theta,V) \\ \ell(\theta,V) &= -\log P(x_1,\ldots,x_n;\theta,V) \\ &= \sum_{i=1}^n \left(d^2(\mathbf{f}_i,\mathbf{v}_i)/T + \log\left(\sum_{j=1}^n \exp(-d^2(\mathbf{f}_i,\mathbf{v}_j)/T) \right) \right). \end{aligned}$$

Experiments

We performed experiments on unsupervised image classification tasks on MNIST, CIFAR10, SVHN and FashionMNIST dataset. We also evaluate the convergence rate of the algorithms on the image classification task as well as transferability of the learned feature representation.
In addition, we explored using the learned feature representation as initialization for visual based reinforcement learning tasks. We evaluate the learned feature representation on the Pong, Enduro and TORCS environment.

Results

We now present all the results here. First we present the results on image classification task.

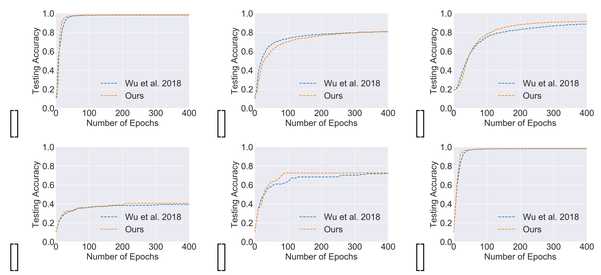

Figure: Results on convergence rate comparing ours with [1] on three datasets: (a) MNIST, (b) CIFAR10, (c) SVHN, (d) MNIST transfer to SVHN, (e) SVHN transfer to MNIST, (f) MNIST transfer to FashionMNIST.

Then we present results on reinforcement learning.

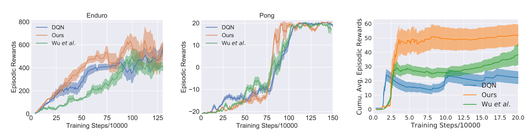

Figure: Reinforcement learning task training reward curve for Enduro, Pong and TORCS environment.

Additional Results

Here are some additional results on image classification.

Table: Comparison of image classification accuracy on various image classification datasets.

Method	MNIST	CIFAR10	SVHN
Ours	**98.51%**	**82.53%**	**91.39%**
[1]	97.85%	80.65%	88.84%

Table: Transfer Learning Results. Every column shows the source dataset, and every row shows the target dataset.

Target\Source	MNIST	SVHN
MNIST (Ours)	-	**72.74%**
SVHN (Ours)	**40.63%**	-
FashionMNIST (Ours)	**98.37%**	-
MNIST [1]	-	71.88%
SVHN [1]	39.41%	-
FashionMNIST [1]	98.21%	-

Conclusion

In this work, we propose a novel manifold hypersphere unsupervised learning method for feature representation learning. We compared with one of the state of the art method in this field and our results show that our method exceeds the performance of [1] in terms of better accuracy, convergence rate and also transferability. Moreover, we show by experiments in reinforcement learning that our learned feature extractor can help to improve sample efficiency in vision based reinforcement learning tasks.

References

[1] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin.
Unsupervised feature learning via non-parametric instance discrimination.
In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.