Deep Variational Metric Learning

GROUP -21 : Lab_Rats

Group Members:
Ujjwal Tiwari (2019701016) - (MS by Research)
Shreyank Jyoti (2018900069) - (MS by Research)
Madan N S (2018900075) - (PGSSP)

Mentored By: Kunal

Github Link

Motivation

In the conventional metric learning methods, intra-class variance and class centers are entangled,which limits to:

• Given a limited dataset with a large range of variance within classes, current metric learning methods are easily over-fitting and lose discriminative power on unseen classes;

• Without disentangling intra-class variance and class centers, current methods learn a metric by exploring the boundary among classes, which means numerous easy negative samples contribute little to the training procedure.

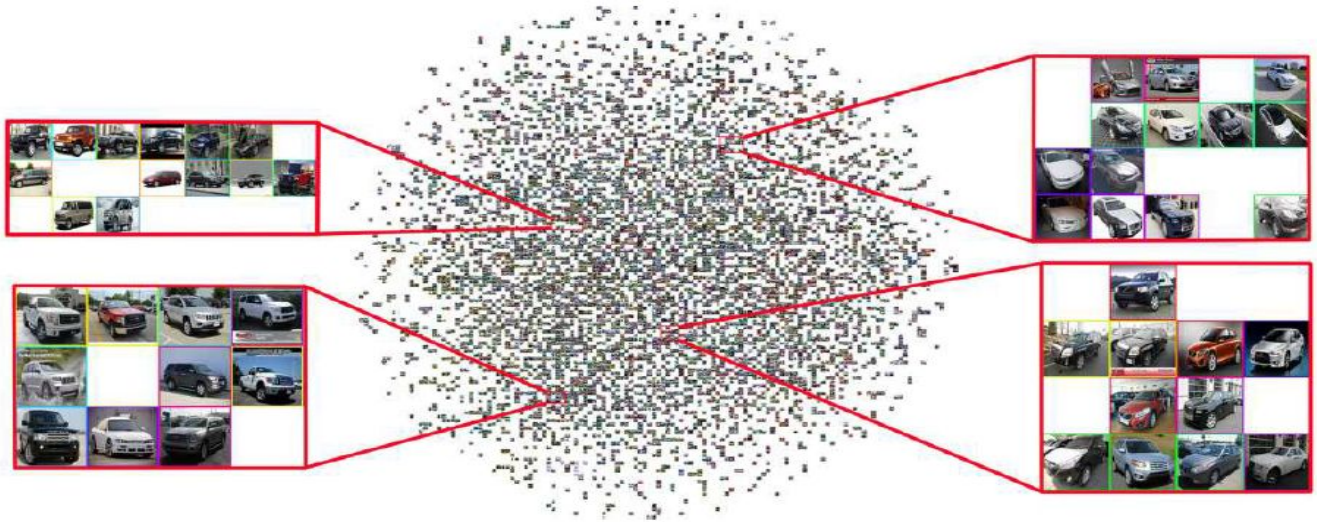

Fig. 1. Our insight: in the central latent space, the distribution of intra-class variance is independent on classes. This is the visualization of central latent space of features learned with the N-pair loss [23] using Barnes-Hut t-SNE [30] on the Cars196 test set. The color of the bounding box for each image represents the class label. Here we construct central latent space through subtracting samples' class centers from their features. We assumed and verified that similar change of original images, like the same pose change or the same view-point change, affects their features in a similar way. (Best viewed when zoomed in.)

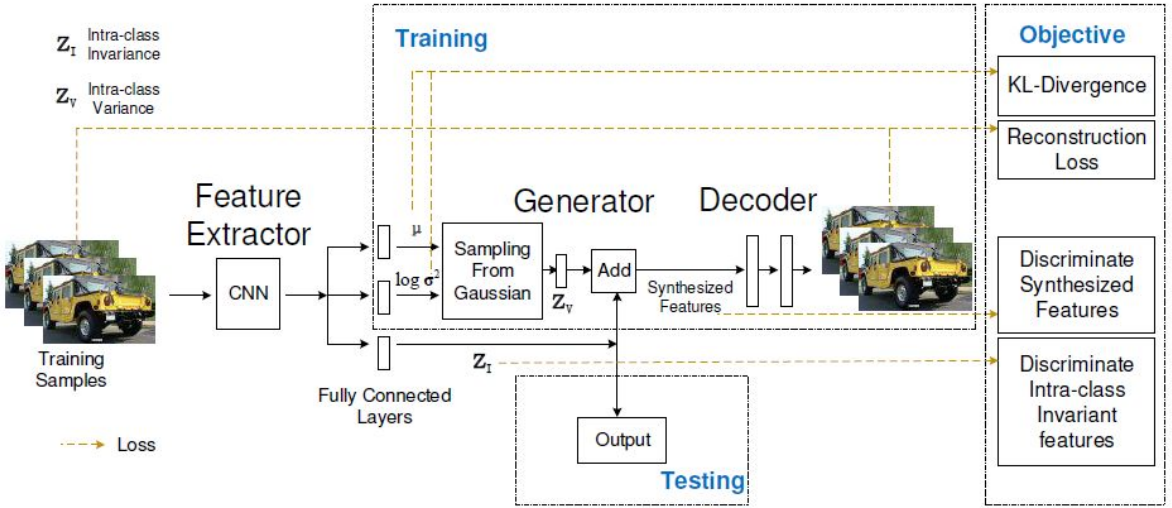

Fig. 2. Our proposed DVML framework. Taking the output of a backbone feature extractor as input, the following layers consist of two parts. The upper part is to model intra-class variance, and it only works in the training procedure. The third fully connected layers following the feature extractor is used to learn intra-class invariant features $\mathbf{z}_I$, namely, the class centers, which is also the output of our model. The generator takes as inputs the class centers $\mathbf{z}_I$ and the features sampled from the learned distribution $\mathcal{N}(\mathbf{z}_V; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\mathbf{I})$, and then outputs element-wise sum of them as synthesized discriminative samples. In order to reduce computation cost, we reconstruct the 1024-dimension features, which are the output of the backbone feature extractor, instead of the whole images, . (Best viewed when zoomed in.)

Methodology

Loss functions:

1) the KL divergence between learned distribution and isotropic multivariate Gaussian.

2) the reconstruction loss of original images and images generated by the decoder

3) the metric learning loss of learned intra-class invariance;

4) the metric learning loss of the combination of sampled intra-class variance and learned intra-class invariance.

Final Loss: Weighted sum of all the losses.

Contribution

Disentangle intra-class variance and intra-class invariance, which makes it possible to explicitly learn appropriate class centers by simultaneously minimizing L4.

Second, different from previous hard negative mining methods which ignore numerous easy negative samples, with the learned distribution of intra-class variance, we generate discriminative samples which contains the possible intra-class variance over the whole training set.

It is obvious that our discriminative sample generation is entirely different from conventional data augmentation methods.

Datasets

• The **CUB-200-2011 dataset** contains 11,788 images from 200 bird species. We take the first 100 classes with 5,864 images for training, and the rest 100 classes with 5,924 images for testing.

• The **Cars196 dataset** contains 16,185 images of 196 car types. We take the first 98 classes with 8,054 images for training, and the rest 98 classes with 8,131 images for testing.

• The **Stanford Online Products dataset** contains 120,053 images of 22,634 products. We take the first 11,318 classes with 59,551 images for training, and the rest 11,316 classes with 60,502 images for testing.

Table 1. Comparisons of clustering and retrieval performance (%) on the Cars196 dataset

Mehtod	NMI	F_1	R@1	R@2	R@4	R@8
Triplet [36,18]	56.2	21.3	58.5	68.8	77.1	84.2
DVML+Triplet	**61.1**	**28.2**	**64.3**	**73.7**	**79.2**	**85.1**
N-pair [23]	62.9	31.9	72.3	79.9	86.8	90.9
DVML+N-pair	**66.0**	**34.6**	**80.4**	**85.8**	**91.8**	**95.1**
Contrastive [7]	44.8	11.2	35.8	47.5	59.7	71.5
Lifted [25]	60.0	27.9	70.0	79.5	86.8	92.0
Angular [33]	61.2	30.8	70.1	80.2	86.7	91.6
$Triplet_2$+DWS [37]	65.4	34.3	78.9	85.6	91.0	94.7
DVML+$Triplet_2$+DWS	**67.6**	**36.8**	**82.0**	**88.4**	**93.3**	**96.3**
HDC [40]	-	-	73.7	83.2	89.5	93.8
Proxy-NCA [16]	64.9	-	73.2	82.4	86.4	88.7

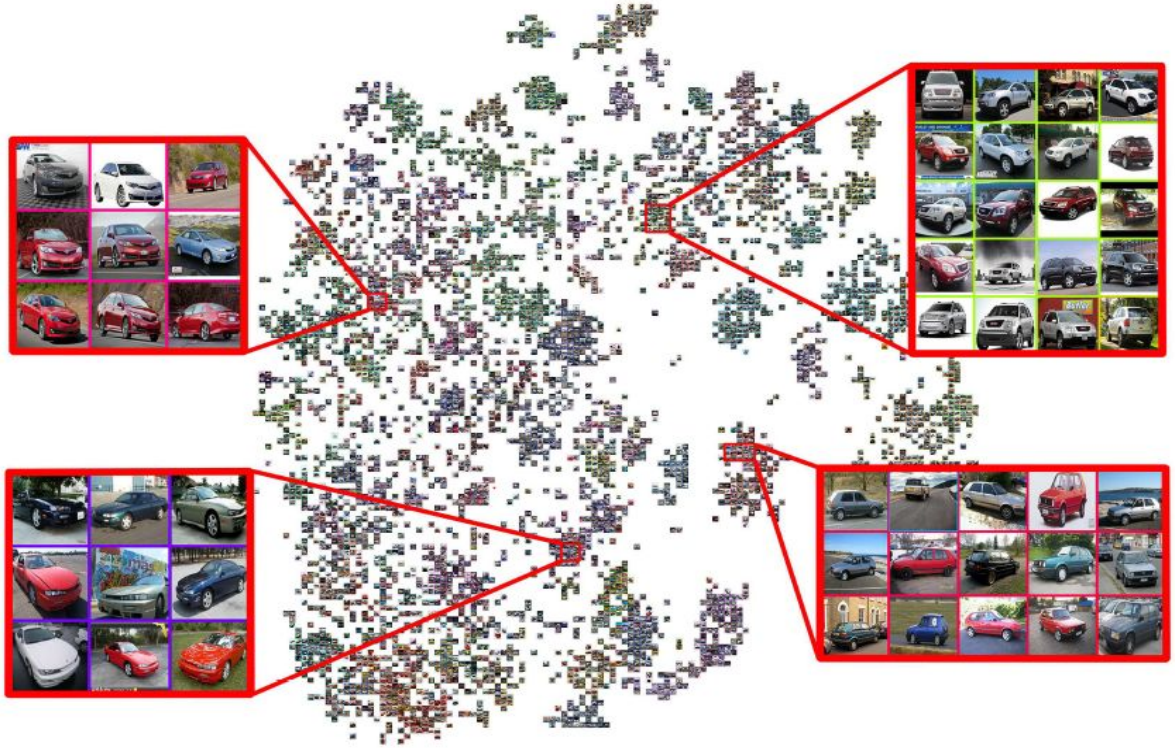

Fig. 4. Visualization of the proposed DVML+N-pair with Barnes-Hut t-SNE [30] on the Cars196 test set. The color of the bounding box for each image represents the label. (Best viewed when zoomed in.)