Exemplar Normalization for Learning Deep Representation

Supplemental Material

Ruimao Zhang[1*], Zhanglin Peng[1*], Lingyun Wu[1], Zhen Li[3,4], Ping Luo[2]
[1] SenseTime Research, [2] The University of Hong Kong,
[3] The Chinese University of Hong Kong (Shenzhen), [4] Shenzhen Research Institute of Big Data
{zhangruimao, pengzhanglin, wulingyun}@sensetime.com, lizhen@cuhk.edu.cn, pluo.lhi@gmail.com

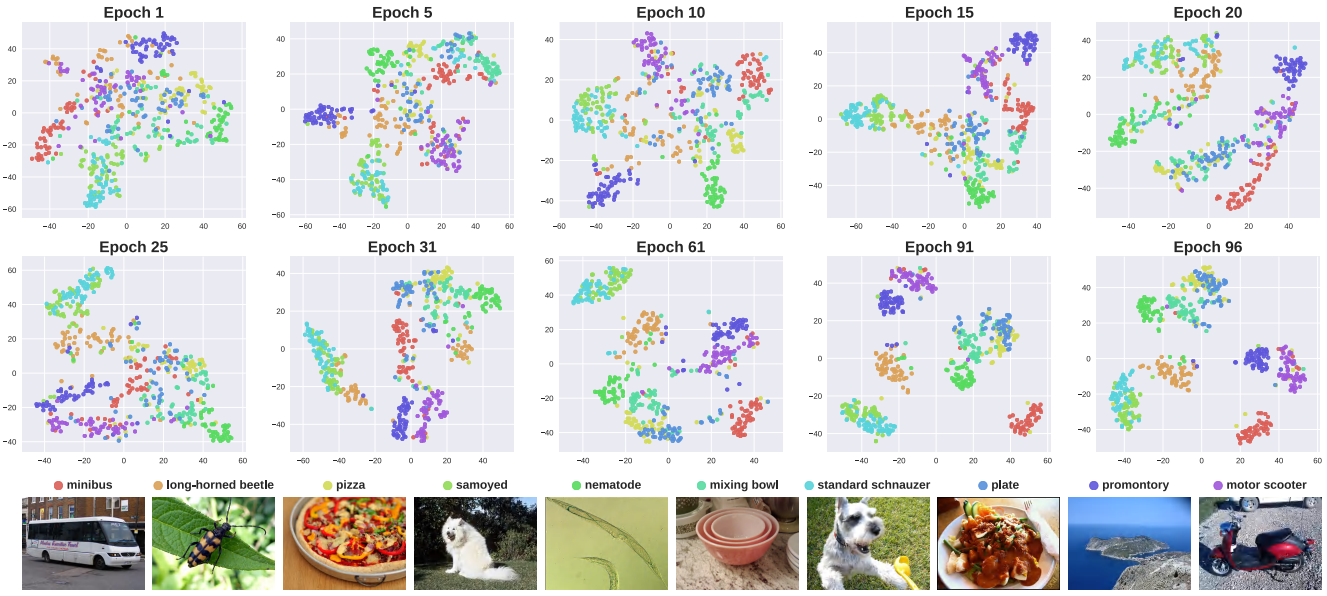

Figure 1. The visualization of sample clustering on 10 categories of ImageNet validation set by using learned important ratios. For each epoch, the concatenated important ratios of each sample are reduced to 2-D by using t-SNE [2]. The corresponding epoch number is shown in the top of each sub-figure. Note that the 31, 61 and 91 epoch are three epochs after the learning rate adjustment. Each color in the figure is corresponding to a specific semantic category. Better view in color with zooming in.

A. Sample Clustering via Important Ratios

Exemplar Normalization (EN) provides another perspective to understand the structure information in CNNs. To further analyze the effect of proposed EN on capturing the semantic information, we concatenate the learned important ratios in all of the EN layers for the input images and adopt t-Distributed Stochastic Neighbor Embedding (t-SNE) [2] to reduce the dimensions to 2-D. The visualization of these samples are shown in Fig. 1.

In practice, we train EN-ResNet50 on the ImageNet [1] training set. The normalizer pool used in EN is $\{$ IN, LN, BN $\}$. Then we randomly select 10 categories from ImageNet validation set to visualize the sample distribution. For each categories, all of the validation samples are used (*i.e.* 50 samples per category). The name of the selected categories and related exemplary images are present at bottom of Fig. 1. To visualize each sample, we extract and concatenate its important ratios from all of the EN layer in EN-ResNet50. Thus the dimension of concatenated important ratios is $53 \times 3 = 159$. Then we use the open source of t-SNE[1] to reduce the dimension from 159 to 2 to visualize the sample distribution. We select 10 typical training epochs to show the clustering dynamic in the training phase.

According to Fig. 1, we have the following observations. (1) The learned important ratios can be treated as one type of structure information to realize **semantic preservation**.

[1] https://lvdmaaten.github.io/tsne/

When the model converges, *i.e*. at 96 epoch, the samples with the same label are grouped into the same cluster. It further demonstrates different categories tend to select different normalizers to further improve their representation abilities, as well as the prediction accuracy of the model. (2) The learned important ratios in EN also makes **appearance embedding** possible. For example, the `samoyed` and `standard schnauzer` have the same father category according to the WordNet[2] hierarchy and the samples in these two categories share the same appearance. Thus, the distance between the corresponding two clusters are small. The same result also achieves in category `pizza` and `plate`. But cluster `samoyed` is far away from cluster `pizza` since they provide great difference in appearance. (3) We also investigate the **clustering dynamic** in Fig. 1. We show the sample distributions in 10 different epochs of training process. In the beginning of the model training, all of the samples are uniform distributed and none of semantic clusters are generated. From 5 epoch to 25 epoch, the semantic clusters are generated rapidly along with the model optimization. The semantic clusters are basically formed after 31 epoch, which is the first epoch after the first time to decay the learning rate. After that, the sample distribution are slightly adjusted in the rest epochs.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[2] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[2] `https://wordnet.princeton.edu/`