Deep metric learning using Triplet network

{ Elad Hoffer, Nir Ailon } Technion Israel Institute of Technology, Department of Computer Science

Objectives

Deep learning has proven itself as a successful set of models for learning useful semantic representations of data. These, however, are mostly implicitly learned as part of a classification task. In this work, we aim to learn useful representations by distance comparisons.
Accordingly, we try to fit a metric embedding S so that:

$$S(x,x_1) > S(x,x_2), \quad \forall x,x_1,x_2 \in \mathbb{P} \quad \text{for which } r(x,x_1) > r(x,x_2).$$

where $r(x,x')$ is a rough similarity measure given by an oracle.
We focus on finding an L_2 embedding, by learning a function $F(x)$ for which $S(x,x') = \|F(x) - F(x')\|_2$.

Introduction - Triplet Network

A *Triplet network* is comprised of 3 instances of the same feed-forward network (with shared parameters). When fed with 3 samples, the network outputs 2 intermediate values - the L_2 distances between the embedded representation of two of its inputs from the representation of the third.

$$TripletNet(x,x^-,x^+) = \begin{bmatrix} \|Net(x) - Net(x^-)\|_2 \\ \|Net(x) - Net(x^+)\|_2 \end{bmatrix} \in \mathbb{R}_+^2 .$$

In words, this encodes the pair of distances between each of x^+ and x^- against the *reference* x.

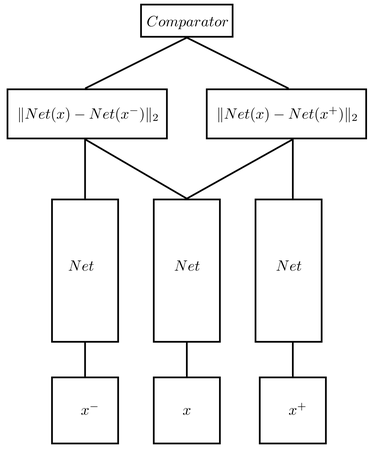

Figure 1: Triplet network structure

Metric learning using deep convolutional embedding network

Training was preformed by feeding a *Convolutional* embedding network with samples where, x and x^+ are of the same class, and x^- is of different class. In order to output a comparison operator from the model, a SoftMax function is applied on both outputs - effectively creating a ratio measure, so that

$$Loss(d_+,d_-) \to 0 \quad \text{iff} \quad \frac{\|Net(x) - Net(x^+)\|}{\|Net(x) - Net(x^-)\|} \to 0$$

By using the same shared parameters network, we allow the back-propagation algorithm to update the model with regard to all three samples simultaneously.

Results - Achieving useful semantic feature embedding

We experimented with 4 datasets *CIFAR10, SVHN, MNIST, STL10*. We can see a significant clustering by semantic meaning, confirming that the network is useful in embedding images into the Euclidean space according to their content. Similarity between objects can be easily found by measuring the distance between their embedding and, as shown in the results, can reach high classification accuracy using a simple subsequent linear classifier.

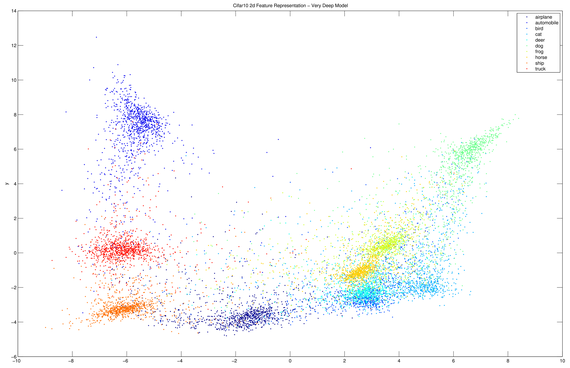

Figure 2: CIFAR10 - Euclidean representation

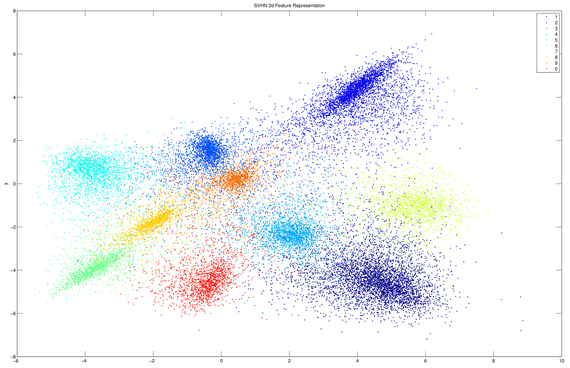

Figure 3: SVHN - Euclidean representation

Using learned representations for classification

These results are comparable to state-of-the-art results with a deep learning model trained explicitly to classify samples, without using any data augmentation [a].
We note that similar results are achieved when the embedded representations are classified using a linear SVM model or KNN classification with up to 0.5% deviance from the results.
We were also able to show that the Triplet Network provided better results than its immediate competitor, the Siamese network, which was trained using a contrastive loss and the same embedding network.

[a] Using standard transformations to artificially enhance the training set

Dataset	TripletNet	SiameseNet	Best known result
Mnist	99.54±0.08%	97.9±0.1%	99.61%
Cifar10	87.1%	-	90.22%
SVHN	95.37%	-	98.18%
STL10	70.67%	-	67.9%

Figure 4: Classification accuracy (no data augmentation)

Future Research

As the Triplet net model allows learning by comparisons of samples instead of direct data labels, usage as an unsupervised learning model is possible. Future investigations can be performed in several scenarios:

- **Using spatial information.** Objects and image patches that are spatially near are also expected to be similar from a semantic perspective. Therefore, we could use geometric distance between patches of the same image as a rough similarity oracle $r(x,x')$, in an unsupervised setting.
- **Using temporal information.** The same is applicable to the time domain, where two consecutive video frames are expected to describe the same object, while a frame taken 10 minutes later is less likely to do so. Our Triplet net may provide a better embedding and improve on past attempts.

It is also well known that humans tend to be better at accurately providing comparative labels. Our framework can be used in a crowd sourcing learning environment. Furthermore, it may be easier to collect data trainable on a Triplet network, as comparisons over similarity measures are much easier to attain.

Conclusions

This work provides evidence that the representations that were learned are useful for classification in a way that is comparable with a network that was trained explicitly to classify samples.
We have also shown how this model learns using only comparative measures instead of labels, which we can use in the future to leverage new data sources for which clear out labels are not known or do not make sense (e.g hierarchical labels).

References

[1] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.

[2] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 2005.

[3] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. *The Journal of Machine Learning Research*, 2010.