DEEP REPRESENTATION LEARNING USING TRIPLET NETWORK

Elad Hoffer
Department of Computer Science
Technion, Israel
ehoffer@tx.technion.ac.il

Nir Ailon
Department of Computer Science
Technion, Israel

ABSTRACT

The abstract paragraph should be indented 1/2 inch (3 picas) on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word ABSTRACT must be centered, in small caps, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 TRIPLET NETWORK

Triplet network (inspired by "Siamese network") is comprised of 3 instances of the same feed-forward network (with shared parameters).
When fed with 3 samples, the network outputs 2 values - the L_2 distance between the embedded representation of 2 input from the representation of the third.
If we will denote the 3 inputs as x y_1 and y_2, and the embedded representation of the network as $Net(x)$, the output will be the vector

$$TripletNet(x, y_1, y_2) = \begin{bmatrix} \|Net(x) - Net(y_1)\|_2 \\ \|Net(x) - Net(y_2)\|_2 \end{bmatrix}$$

1.1 TRAINING

Training is preformed by feeding the network with random samples, where x and y^+ are of the same class, and y^- is of different class. The objective is to classify correctly which sample is of the same class as x. Two same-class samples should have a lower L_2 distance after the network embedding. In order to distinguish between the highest distance, a SoftMax function is applied on both outputs - effectively creating a ratio measure.
By using the same shared-parameters network, we allow the back-propagation algorithm to update the model with regard to all samples *simultaneously*. Training is done by simple stochastic-gradient-descent on a negative-log-likelihood loss with regard to the 2-class problem.

2 CITATIONS, FIGURES, TABLES, REFERENCES

These instructions apply to everyone, regardless of the formatter being used.

2.1 CITATIONS WITHIN THE TEXT

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis (as in "See Hinton

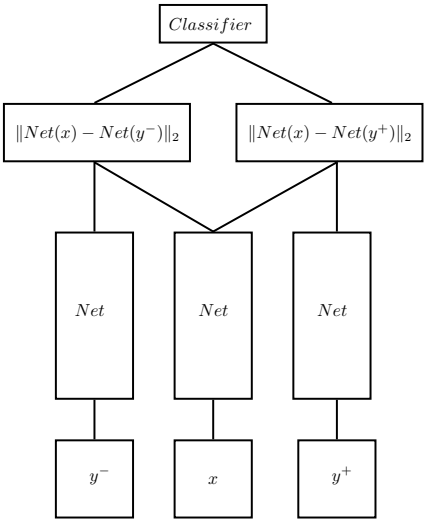

Figure 1: Triplet network structure

et al. (2006) for more information."). Otherwise, the citation should be in parenthesis (as in "Deep learning shows promise to make progress towards AI (Bengio & LeCun, 2007).").

The corresponding references are to be listed in alphabetical order of authors, in the REFERENCES section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

2.2 FOOTNOTES

Indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[2]

2.3 TABLES

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table **??**.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

3 TESTS AND RESULTS

3.1 DATASET

The Triplet Network was trained on the *Cifar10* dataset which consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images.
Each training instance is a uniformly sampled 3 images, 2 of which are of the same class, and the third is of a different class. Each training epoch consists of 640,000 such instances (randomly chosen each epoch), and a fixed 64,000 instances used for test (from the test images).

[1] Sample of the first footnote
[2] Sample of the second footnote

3.2 EMBEDDING NET

The embedding net used is a convolutional-network, consisting of 3 convolutional and max-pooling layers. The network outputs a fixed vector of length 128 as embedded representation. These values are constrained to be positive by a *ReLU* non-linearity ($ReLU(x) = max\{0, x\}$). The network also uses the Dropout regularization technique to avoid over-fitting.

3.3 RESULTS

After training for 30 epochs, the network reached a fixed error of 6% (error of triplet comparisons). By using the representation of the data gained from the embedding network, an accuracy of 84% was reached on all 10 classes. To achieve this, the representations were simply fed into a multi-class SVM model.
This results is comparable to state-of-the-art results with out any data augmentation.
Another side-affect noticed, is that the representation seems to be sparse - about 25% non-zero values. This is very helpful when used in later classification.

3.4 2D VISUALIZATION OF CLASSES

By using PCA, a transformation of the representation into 2d can be displayed. 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 (figure 2).

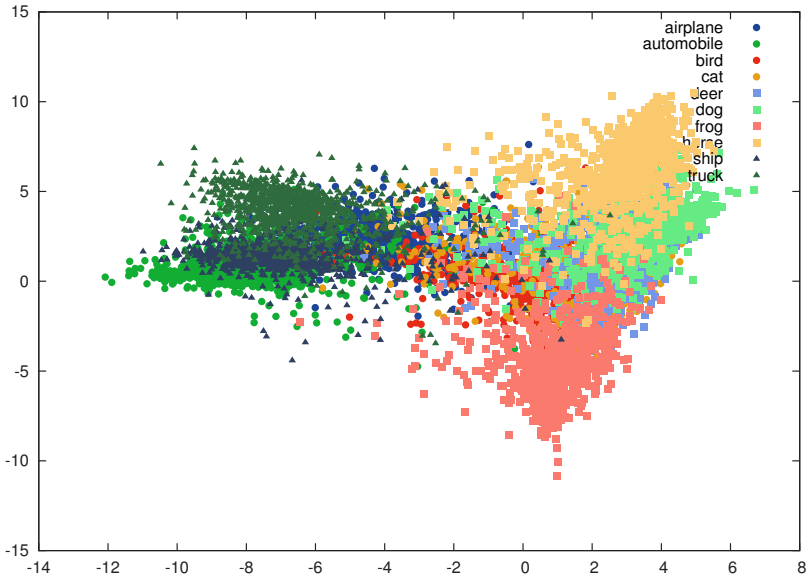

Figure 2: Euclidean 2d Representation

4 UNSUPERVISED LEARNING

We can use the model for unsupervised training, by using patches from the same image (with different scale, augmentation) as x and y^+, and an image taken from a different source as y^-. This seems to behave in a desirable way (although of course much less discrminative than the supervised version) as can be seen in Figure (3).

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

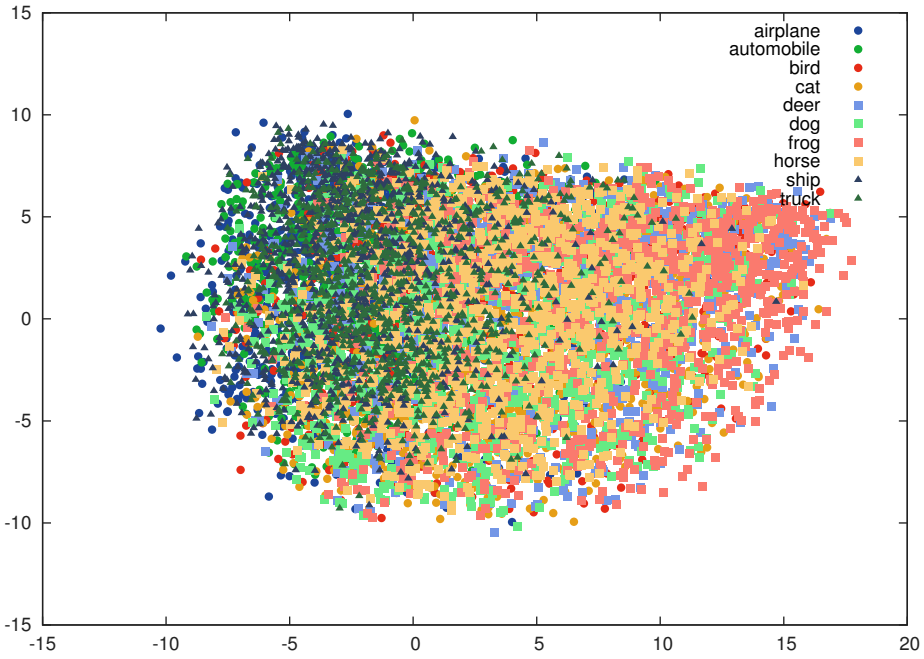

Figure 3: Euclidean 2d of representation in unsupervised learning

REFERENCES

Bengio, Yoshua and LeCun, Yann. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Hinton, Geoffrey E., Osindero, Simon, and Teh, Yee Whye. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.