Deep Learning using contrastive measures

Elad Hoffer advised by Prof. Nir Ailon

Technion Israel Institute Of Technology

Background

Deep learning of representations

Deep learning proved itself over the last few years as a very useful learning technique, often yielding state of the art results in many real-world problems.

- These models were found to supply representations of data that are useful for other tasks in the domain.
- Those "feature representations", although observed to be pivotal to the success of deep models, are usually not explicitly learned, and are often a by-product of a classification task.

Learning by comparisons

Taking inspiration from human learning, we often find that humans tend to relate more easily to comparative measures than to explicit discrimination.

- E.g: Instead of answering the question:

What is this flower? / What are the key properties of it?

Learning by comparisons

Taking inspiration from human learning, we often find that humans tend to relate more easily to comparative measures than to explicit discrimination.

- E.g: Instead of answering the question:

What is this flower? / What are the key properties of it?

- It is much easier for humans to answer

Which of these is more similar? / What are the shared properties?

Learning by comparisons

Key questions that arise:

- Can machine learning models also benefit from these kind of comparisons?
- Can this allow us to use comparative measures that occur naturally in real life scenarios?
  E.g - objects with physical/temporal proximity often share semantic relatedness.

Triplet Network

Triplet network

Deep metric learning using Triplet network - Hoffer, Ailon 15'

Our first attempt at learning by comparison

- Learn feature representations of object by comparing 3 items: 2 of them are expected to belong to one semantic class, while the third belongs to another
- Learning is done by contrasting - the two related object should have a much smaller embedded distance than the third object.
- Different from Siamese network - where the embedded distance is always with respect to predefined margin α

Triplet network

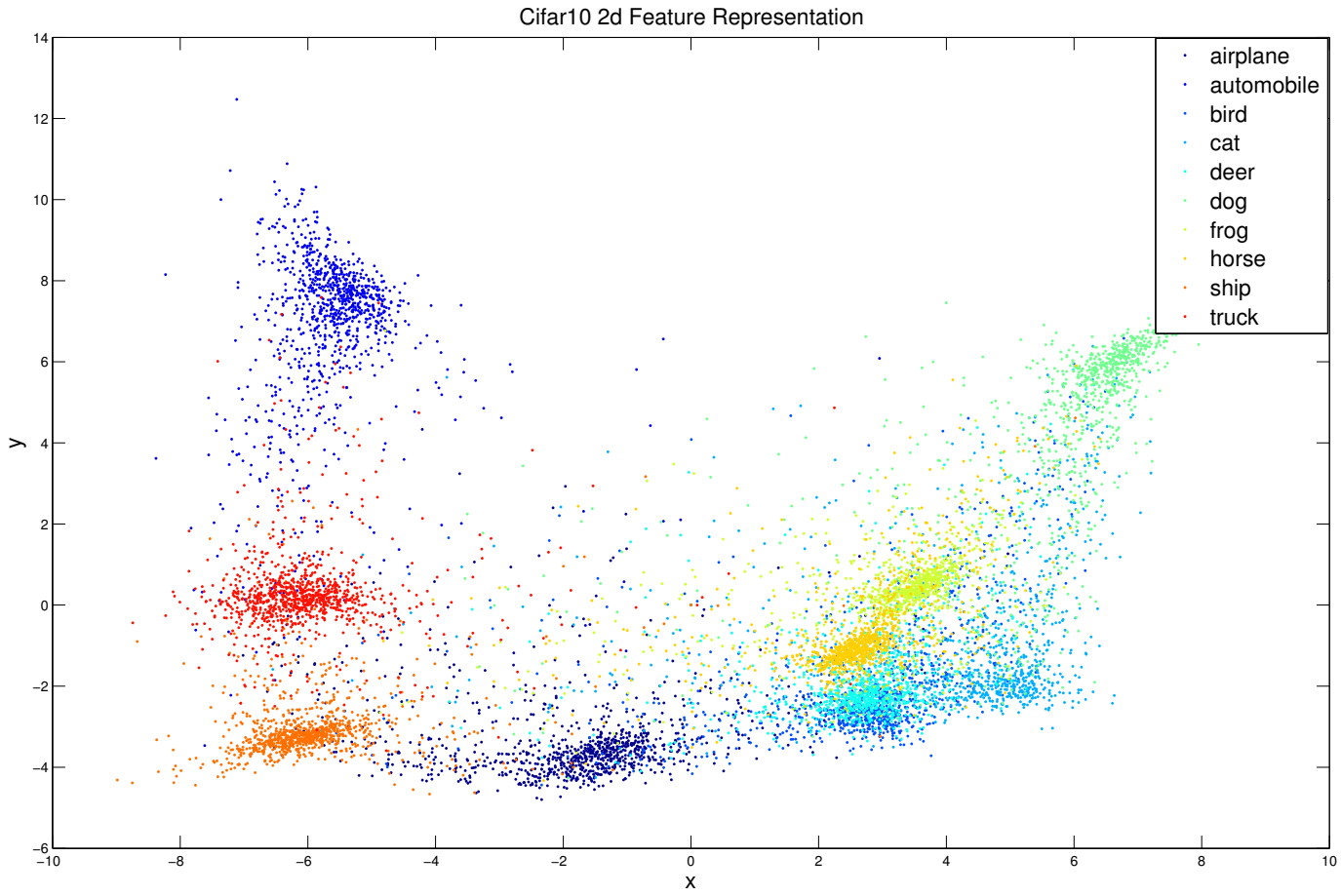
Training is preformed by feeding the network with samples where, as explained above,

- x and x^+ are of the same class, and x^- is of different class.
- Each image is embedded by a convolutional network $Net(x)$
- The network architecture allows the task to be expressed as a 2-class classification problem, where the objective is to correctly classify which of x^+ and x^- is of the same class as x.

$$\mathscr{L}(x, x_+, x_-) = \frac{e^{\|Net(x)-Net(x^-)\|_2}}{e^{\|Net(x)-Net(x^+)\|_2} + e^{\|Net(x)-Net(x^-)\|_2}}$$

Objective and Loss functions

This proved to work very good in practice:



Spatial contrasting

Leveraging comparative measures

We now turn to the second question - can we leverage other comparative indicating signals to learn in unsupervised / semi-supervised fashion?

- Specifically, can the spatial proximity of visual object be useful to build useful semantic representations of them

Spatial contrasting

We look at images patches $\tilde{x}^{(m)}$ taken from an image x.

- A convolutional network extracts spatial features f so that $f^{(m)} = F(\tilde{x}^{(m)})$.
- We wish to build a model such that for two features representing patches taken from the same image $\tilde{x}_i^{(1)}, \tilde{x}_i^{(2)} \in x_i$ for which $f_i^{(1)} = F(\tilde{x}_i^{(1)})$ and $f_i^{(2)} = F(\tilde{x}_i^{(2)})$, the conditional probability $p(f_i^{(1)}|f_i^{(2)})$ will be maximized.
- Conversely, we want our model to minimize $p(f_i|f_j)$ for i, j being two patches taken from distinct images.

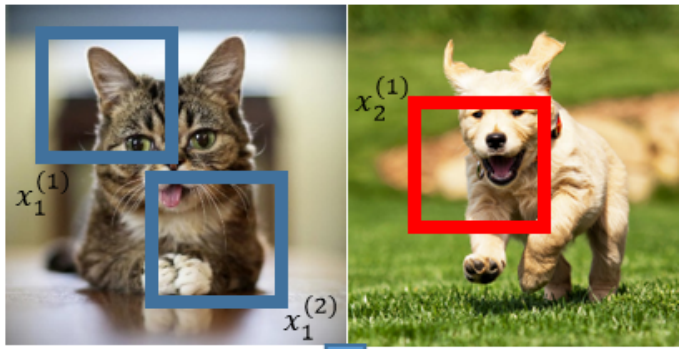
Spatial contrasting

Following the logic presented before, we will need to sample *contrasting patch* $\tilde{x}_j^{(1)}$ from a different image x_j such that $p(f_i^{(1)}|f_i^{(2)}) > p(f_j^{(1)}|f_i^{(2)})$.

- We will use a distance ratio to represent the probability two feature vectors were taken from the same image.
- The resulting training loss for a pair of images will be defined as

$$\mathcal{L}_{SC}(x_1, x_2) = -\log \frac{e^{-\|f_1^{(1)} - f_1^{(2)}\|_2}}{e^{-\|f_1^{(1)} - f_1^{(2)}\|_2} + e^{-\|f_1^{(1)} - f_2^{(1)}\|_2}}$$

Spatial contrasting



ConvNet

$f_1^{(2)}$ $f_1^{(1)}$ $f_2^{(1)}$

Spatial Contrasting

$$\min\left\{-log\frac{e^{-\left\|f_1^{(1)}-f_1^{(2)}\right\|_2}}{e^{-\left\|f_1^{(1)}-f_1^{(2)}\right\|_2}+e^{-\left\|f_1^{(1)}-f_2^{(1)}\right\|_2}}\right\}$$

Spatial contrasting

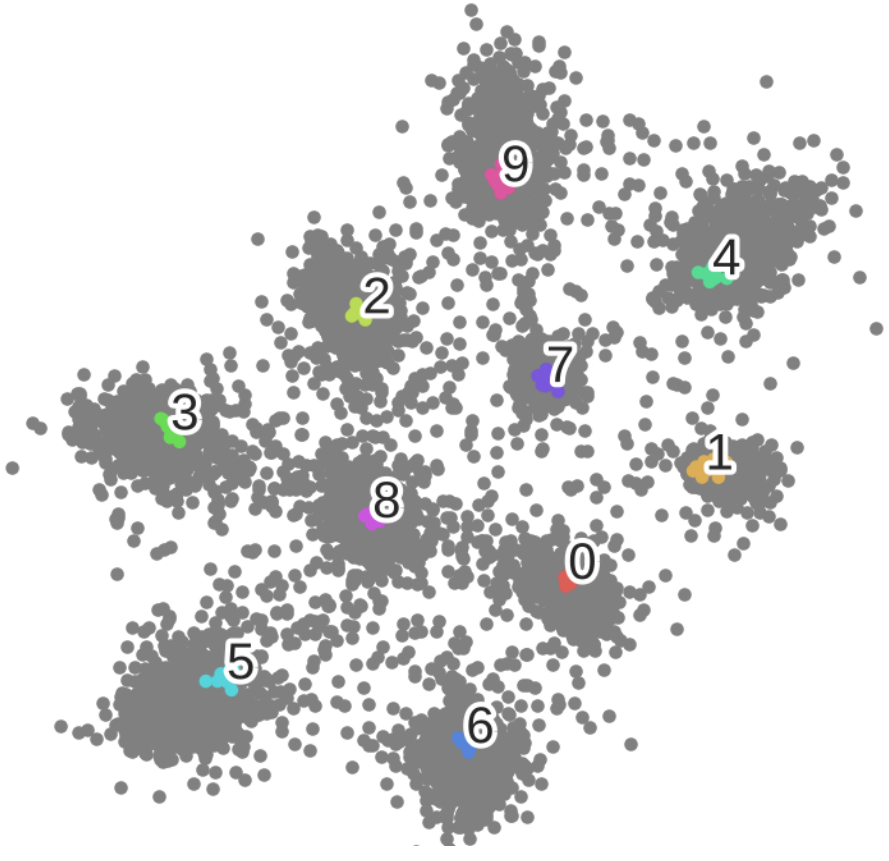
Training this way can provide a technique for unsupervised learning using unlabeled image data. This proved to allow state-of-the-art results for STL10 - dataset where most of the data is unlabeled.

Model	**STL-10 test accuracy**
Triplet network	70.7%
Exemplar Convnets	72.8%
Target Coding	73.15%
Stacked what-where AE	74.33%
Spatial contrasting initialization	81.34% ± 0.1
The same model without initialization	72.6% ± 0.1

Semi-supervised learning by contrasting

Semi-supervised

Focusing on a semi-supervised setting - where only a small subset of examples is labeled, we try to learn a metric embedding that forms clusters around the labeled examples.



Semi-supervised learning by contrasting

We will define a discrete distribution for the embedded distance between a sample $x \in \mathcal{X}$, and c labeled examples $z_1, ..., z_c \in X_L$ each belonging to a different class:

$$P(x; z_1, ..., z_c)_i = \frac{e^{-\|F(x)-F(z_i)\|^2}}{\sum_{j=1}^{c} e^{-\|F(x)-F(z_j)\|^2}}, \quad i \in \{1...c\}$$

This definition assigns a probability $P(x; z_1, ..., z_c)_i$ for sample x to be classified into class i, under a 1-nn classification rule, when $z_1, ..., z_c$ neighbors are given.

Semi-supervised learning by contrasting

We can now state our objectives as entropy measures over the defined distribution with respect to labeled samples $z_1, ..., z_c$:

- Given labeled example x_l, we will minimize the cross-entropy between class-indicator $I(x_l)$ and the distance-distribution:

$$\mathscr{L}_L(x_l, z_1, ..., z_c) = H\left(I(x_l), P(x_l; z_1, ..., z_c)\right)$$

- Given unlabeled example x_u, we will minimize the entropy of the underlying distance distribution:

$$\mathscr{L}_U(x_u, z_1, ..., z_c)_U = H(P(x_u; z_1, ..., z_c))$$

Semi-supervised

Learning using this method, we can achieve very good classification accuracy, with only a small subset of labeled exampels.

For example, learning MNIST with only 100 labeled samples.

Model	Test error %
EmbedCNN	7.75
SWWAE	9.17
Ladder network	0.89 (± 0.50)
Conv-CatGAN	1.39 (± 0.28)
Ours	**0.78 (± 0.3)**

Future work

Further explorations using comparative measures are

- Can we better sample our contrast examples?
- Can we do the same for temporal data? Using time adjacency as a comparison criterion

Acknowledgements

The research leading to these results has received funding from the European Research Council under European Union's Horizon 2020 Program, ERC Grant agreement no. 682203 "SpeedInfTradeoff".