Master Thesis Proposal: Unsupervised Representation Learning for Object Detection and Classification

Candidate: Fedor Chervinskii
Advisor: Victor Lempitsky

December 4, 2015

1 Background

In the area of computer vision, as well as in many other areas of data analysis, most of problems can be formalized as object(pattern) detection and/or classification. Common approach in the field is to train a classifier that would predict the targeted values(classes) or which would predict existence of a targeted pattern/object using a reasonable amount of labeled exemplars. This approach is known as supervised learning. However, while raw data mostly is much easier accessible, that is not the case for labeled data. In some cases collecting necessary labels for the data becomes the most difficult part of a problem, especially, in terms of human resources/time or/and money.

On the other side, human brain solves recognition and detection tasks reliably even based on a small number of references, sometimes one exemplar is enough ("one shot learning"). This ability is an object of current research interest both in machine learning and in computational neuroscience. General intuition about it is based on assumptions that 1) human brain is sharing world's representations across a large number of tasks, that leads to knowledge generalisation 2) it can learn effective representation of an input stimuli based just on prior distributions. The goal of this research project is to use formalizations of these qualities in order to build a framework and suggest a pipeline that combine most advanced deep learning methods and techniques in the way that allows to solve different tasks with minimal supervision.

2 Problem Statement

Suppose we have the data X and some targeted value $f(x)$ that we want to predict for this data. For a computer vision task X would be a set of images/videos and an f values - objects/classes that should be detected/predicted. For some amount of the given data we can provide groundtruth values $f(x) = y$ of a target. We will assume that amount of raw data is always significantly bigger than we can provide labels. We want to find an algorithm that would predict target value using information contained both in labeled and unlabeled parts of the dataset. We want this algorithm to be optimal in terms of labeled/unlabeled

data amounts ratio. We also need a empirical/theoretical estimate for a dependence of the performance on this ratio.

3 Objectives

The main objective of the project is a practical realization of an algorithm and it's application to at least two real-world problems. Some tasks, that I've already started to test the defined approach, include cells detection on microscopic images, faces and scene recognition.

Complementary goals of this work would be

- Report of an achieved performance on any of tasks in format of publication or technical report.
- Code publication for public domain.
- Development of a theory supporting the results.

4 Related Work

The work is going to be based on deep learning approaches proposed during last few years. [Bengio et al., 2012] discuss necessary features of a good representation how they could be achieved. [Kingma and Welling, 2013] defined a deep learning architecture named Variational AutoEncoder (VAE) that learns a statistical inference of a data. Experiments show that VAE learns meaningful and interpretable representation of an unlabeled data. Certain improvements in this architecture has been proposed, including Importance Weighted Autoencoders [Burda et al., 2015], Adversarial Autoencoders [Makhzani et al., 2015] and Ladder Networks [Rasmus et al., 2015]. [Kulkarni et al., 2015] shows how we can insert the target value in the vector of latent variables to learned conditioned representations.

5 Methodology

During the work process I am going to use datasets that are publicly available (CIFAR, FacesInTheWild, e.t.c.) together with those provided to Skoltech Computer Vision Group by its' partners. Most of authors listed in Related Work publish the code that they use in their research implemented in one of a few most popular deep learning frameworks. Those are `Theano(Python)`, `Torch(Lua)`, `Matconvnet(Matlab)` and `Caffe`. All of them are open-source and have a flexibility sufficient for implementation of almost any architecture. Main programming language of the project still is a subject of a discussion. Computational time-consuming neural networks learning process can be speeded up by using GPU instances.

Performance metrics commonly used in this type of research such as classification accuracy, detection accuracy and recall are going to be used as objectives to be reported and compared with state-of-the-art.

6 Work Plan

Work on the thesis is going to be split into three periods

1. **ISP** Defining an approach, testing different methods on real images of different types. **120 hours**

2. **3rd term** Performing computational experiments on the data, collecting results. **320 hours**

3. **4th term** Results analysis, work on systematization and inference. Work on report and publication of the results. **320 hours**

7 Potential Impacts

Deep Learning recently showed potential to impact human life in many ways. Speech recognition, face recognition, natural language processing, self-driving cars, disease predictions have already come true thanks to development of deep learning methods. In most of real world applications the algorithm have to deal with huge amount of data incoming in real time. Any framework that could make use of almost unlimited raw data to improve model's representation would push the performance further comparing to standard supervised methods.

As a couple of examples, in medical imaging, a researcher could easily label thousands of cells just by marking a few examples. Self-driving car could recognize a new traffic sign after just a couple of its' occurences. In coming years one surely will see much more examples of such applications.

References

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *CoRR*, abs/1206.5538, 2012. URL `http://arxiv.org/abs/1206.5538`.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *CoRR*, abs/1509.00519, 2015. URL `http://arxiv.org/abs/1509.00519`.

D. P Kingma and M. Welling. Auto-encoding variational bayes. *ArXiv e-prints*, 2013. URL `http://arxiv.org/abs/1312.6114`.

Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167, 2015. URL `http://arxiv.org/abs/1503.03167`.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL `http://arxiv.org/abs/1511.05644`.

Antti Rasmus, Harri Valpola, Mikko Honkala, Mathias Berglund, and Tapani Raiko. Semi-supervised learning with ladder network. *CoRR*, abs/1507.02672, 2015. URL `http://arxiv.org/abs/1507.02672`.