Proposition de recherche doctorale

Weakly and semi-supervised learning for image classification

Mots clés :
- Directeur de thèse : Matthieu Cord
- Co-encadrant(s) :
- Unité de recherche : Laboratoire inconnu!
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal : Divers

Résumé du projet de recherche (Langue 1)
In recent years we have witnessed an explosion of successful applications of deep learning including image recognition, speech recognition, automatic translation, self-driving cars, computers that can beat professional Go players, and recommender systems. Because of the abundance of training data, machine learning techniques can now deal with increasingly large and complex inputs (video, sound, speech, text, etc.). As an example, the ImageNet dataset is one of the major object recognition benchmarks in computer vision, it consists of more than 14 million images of more than 20 thousand classes. Computing power is advancing rapidly with massively parallel GPU architectures, changing the nature of machine learning research. In recent years deep learning has surfaced as one of the most effective approaches to exploit the available data and computing power. In deep learning data is used to train a model that consists of multiple non-linear processing layers — all the way from the raw input data to the target predictions — each of which has trainable parameters. Deep networks designed for these tasks have millions, if not billions, of parameters and take enormous resources to train in terms of data, memory and computing power. The purpose is to study the use of non-labeled data — that they do not require human annotations — for joint supervised and unsupervised deep learning. The idea of using unsupervised methods that fits with supervision is not new [Sudderth90; Larochelle08]. The basic criterion for unsupervised learning is the reconstruction (or generation) of the input. For instance, with an auto-encoder for natural images, the decoder will try to reconstruct the original input (with all details) from the internal representation. This approach, however, might not be optimal for tasks such as image classification where pixel-level details are not relevant, since such models should actually be invariant for many low-level imaging conditions. The recently proposed ladder network [Valpola15] modifies the basic auto-encoder reconstruction scheme by not requiring the reconstruction to be performed from the simultaneously input representation. Instead, the decoder also recovers input from being only higher layers of the encoder by adding lateral connections, thus fine details do not need to be propagated to deep representational layers. Based on the same idea, we will explore other schemes to learn hierarchical representations suitable for abstract high-level tasks such as image classification, while also maintaining the capability to reconstruct fine details in the input to ensure compatibility with reconstruction-based loss functions: We propose to combine deep architectures based on scattering operators [Bruna13] with recent approaches dedicated to generate data (images) from deep models [Mahendran15, Gregor15]. The idea behind scattering deep models [Bruna13] is to attempt at linearizing deformations, such that the variations due to some irrelevant aspects are discarded for the high level representations. This makes, however, the reconstruction of lower levels impossible from higher representations alone. We propose here to explicitly model the invariant and variant features for each training example, such that the reconstruction using deep generation model can be performed with this extended information. We will consider as a baseline the deep top-down RBM nets [Goh13] recently proposed as unsupervised/supervised scheme. Our proposition aims at extending this baseline especially by defining a training scheme where the supervised and unsupervised criterion are learned in a joint manner, and by explicitly modeling the decomposition of variant and invariant representations. One assumption we want to validate is that this explicit decompositions drives the learning towards more effective (robust) representations. The second approach we propose is to explore representations of different nature, and particularly mixed representations composed of both continuous and discrete values as was recently proposed in the supervised case [Tang13, Zhu14]. Mixing these two types of representations can be seen as learning an (image) representation composed of both a concept (discrete values) and context (real values). The concept could be for example a binary vector corresponding to a category (i.e. a cat), while the context will be mapped to a continuous one corresponding to the visual context of the category (i.e. on a red sofa). While this information will be extracted through unsupervised techniques such as auto-encoders, a classifier can make use of the binary information only. Such a model will therefore learn to obtain the information from the binary part of the extracted representation. Since the binary part is a discrete information that will be sampled from the observation, learning techniques such as policy gradient and variational bounds will be developed. The binary representation will have interesting properties in term of computation speed, size, etc. allowing to develop fast and economical algorithms. Our last investigation concerns deep generative models. While deep (convolutional) neural networks have now been used with some success to define very powerful conditional random field (CRF) models, training these models requires supervised datasets. More importantly, these models yield unimodal predictive models. While this is fine for conditional prediction problems (there is typically one desired answer per input), for unsupervised generative models multi-modality is required to have non-trivial distributions. To obtain structured models in an unsupervised manner that allows for multi-modal distributions one can use a (convolutional) deep neural network to transform samples from a simple noise source into samples from a complex generative distribution. Adversarial networks [Goodfellow13] and Variational Bayesian auto-encoders [Kingma14] are examples of methods to train such unsupervised models. Such generative unsupervised models are often tested for the task of image synthesis, and recently with surprising success [Radford16]. We will consider how such methods can be used for semi-supervised learning for semantic segmentation tasks, where only for a part of the images supervised label information is available. The question is to share parameters between the generative and the conditional network. Moreover, we will also consider the application of deep generative models for inpainting tasks, where a missing part of an image has to be generated by the model, conditioned on the available part of the image. While this task can be trained in a supervised manner, it does require complex multi-modal output distributions. ([References]) S. Sudderth and Y. Kergosien. Rule-injection hints as a means of improving network performance and learning time. In EURASIP Workshop on Neural Networks, 1990. H. Larochelle and Y. Bengio. Classification using discriminative restricted boltzmann machines. In ICML, 2008. H. Valpola. From neural PCA to deep unsupervised learning. In Advances in Independent Component Analysis and Learning Machines, chapter 8, pages 143–171. 2015. J. Bruna and S. Mallat. Invariant scattering convolution networks. PAMI, 35(8):1872–1886, 2013. A. Mahendran and A. Vedaldi. Understanding deep image representations by inverting them. In CVPR, 2015. K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra. DRAW: A recurrent neural network for image generation. In ICML, 2015. H. Goh, N. Thome, M. Cord, and J.-H. Lim. Top-down regularization of deep belief networks. In NIPS, 2013. Y. Tang and R. Salakhutdinov. Learning stochastic feedforward neural networks. In NIPS, 2013. Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In NIPS, 2014. I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio. Maxout networks. In ICML, 2013. D. Kingma and M. Welling. Auto-encoding variational Bayes. In ICLR, 2014. A. Radford, L. Metz, and S. Chintala. Presentation learning with deep convolutional generative adversarial networks. In ICLR, 2016.