Abstract

In recent years, deep neural networks have achieved extraordinary performance on supervised learning tasks. Convolutional neural networks (CNN) have vastly improved the state of the art for most computer vision tasks including object recognition and segmentation. However, their success relies on the presence of a large amount of labeled data. In contrast, relatively fewer work has been done in deep learning to handle scenarios when access to ground truth is limited, partial or completely absent. In this thesis, we propose models to handle challenging problems with limited labeled information.

Our first contribution is a neural architecture that allows for the extraction of infinitely many features from an object while allowing for tractable inference. This is achieved by using the `kernel trick', that is, we express the inner product in the infinite dimensional feature space as a kernel. The kernel can either be computed exactly for single layer feedforward networks, or approximated by an iterative algorithm for deep convolutional networks. The corresponding models are referred to as stretched deep networks (SDN). We show that when the amount of training data is limited, SDNs with random weights drastically outperform fully supervised CNNs with similar architectures.

While SDNs perform reasonably well for classification with limited labeled data, they can not utilize unlabeled data which is often much easier to obtain. A common approach to utilize unlabeled data is to couple the classifier with an autoencoder (or its variants) thereby minimizing reconstruction error in addition to the classification error. We discuss the limitations of decoder based architectures and propose a model that allows for the utilization of unlabeled data without the need of a decoder. This is achieved by jointly modeling the distribution of data and latent features in a manner that explicitly assigns zero probability to unobserved data. The joint probability of the data and the latent features is maximized using a two-step EM-like procedure. Depending on the task, we allow the latent features to be one-hot or real-valued vectors and define a suitable prior on the features. For instance, one-hot features correspond to class labels and are directly used for the unsupervised and semi-supervised classification tasks. For real-valued features, we use hierarchical Bayesian models as priors over the latent features. Hence, the proposed model, which we refer to as discriminative encoder (or DisCoder), is

flexible in the type of latent features that it can capture. The proposed model achieves state-of-the-art performance on several challenging datasets.

Having addressed the problem of utilizing unlabeled data for classification, we move to a domain where obtaining labels is a lot more expensive, that is, semantic segmentation of images. Explicitly labeling each pixel of an image with the object that the pixel belongs to, is an expensive operation, in terms of time as well as effort? Currently, only a few classes of images have been densely (pixel-level) labeled. Even among these classes, only a few images per class have pixel-level supervision. Models that rely on densely-labeled images, cannot utilize a much larger set of weakly annotated images available on the web. Moreover, these models cannot learn the segmentation masks for new classes, where there is no densely labeled data.

Hence, we propose a model for utilizing weakly-labeled data for semantic segmentation of images. This is achieved by generating fake labels for each image, while simultaneously forcing the output of the CNN to satisfy the mean-field constraints imposed by a conditional random field. We show that one can enforce the CRF constraints by forcing the distribution at each pixel to be close to the distribution of its neighbors. The proposed model is very fast to train and achieves state-of-the-art performance on the popular VOC-2012 dataset for the task of weakly supervised semantic segmentation of images.