Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Integrated inference and learning of neural factors in structural support vector machines

Rein Houthooft*, Filip De Turck

Ghent University - iMinds, Department of Information Technology, Technologiepark 15, Ghent B-9052, Belgium

ARTICLE INFO

Article history:
Received 30 August 2015
Received in revised form
8 March 2016
Accepted 9 March 2016

Keywords:
Structural support vector machine
Neural factors
Structured prediction
Neural networks
Image segmentation

ABSTRACT

Tackling pattern recognition problems in areas such as computer vision, bioinformatics, speech or text recognition is often done best by taking into account task-specific statistical relations between output variables. In structured prediction, this internal structure is used to predict multiple outputs simultaneously, leading to more accurate and coherent predictions. Structural support vector machines (SSVMs) are nonprobabilistic models that optimize a joint input–output function through margin-based learning. Because SSVMs generally disregard the interplay between unary and interaction factors during the training phase, final parameters are suboptimal. Moreover, its factors are often restricted to linear combinations of input features, limiting its generalization power. To improve prediction accuracy, this paper proposes: (i) joint inference and learning by integration of back-propagation and loss-augmented inference in SSVM subgradient descent; (ii) extending SSVM factors to neural networks that form highly nonlinear functions of input features. Image segmentation benchmark results demonstrate improvements over conventional SSVM training methods in terms of accuracy, highlighting the feasibility of end-to-end SSVM training with neural factors.

1. Introduction

In traditional machine learning, the output consists of a single scalar, whereas in structured prediction, the output can be arbitrarily structured. These models have proven useful in tasks where output interactions play an important role. Examples are image segmentation, part-of-speech tagging, and optical character recognition, where taking into account contextual cues and predicting all output variables at once is beneficial. A widely used framework is the conditional random field (CRF), which models the statistical conditional dependencies between input and output variables, as well as between output variables mutually. However, many tasks only require 'most-likely' predictions, which led to the rise of nonprobabilistic approaches. Rather than optimizing the Bayes' risk, these models minimize a structured loss, allowing the optimization of performance indicators directly [1]. One such model is the structural support vector machine (SSVM) [2] in which a generalization of the hinge loss to multiclass and multilabel prediction is used.

A downside to traditional SSVM training is the bifurcated training approach in which *unary factors* (dependencies of outputs on inputs), and *interaction factors* (mutual output dependencies) are trained sequentially. A unary classification model is optimized, while the interactions are trained post hoc. However, this two-phase approach is suboptimal, because the errors made during the training of the interaction factors cannot be accounted for during training of the unary classifier. Another limitation is that SSVM factors are linear feature combinations, restricting the SSVM's generalization power. We propose to extend these linearities to highly nonlinear functions by means of multilayer neural networks, to which we refer as *neural factors*. Towards this goal, subgradient descent is extended by combining loss-augmented inference with back-propagation of the SSVM objective error into both unary and interaction neural factors. This leads to better generalization and more synergy between both SSVM factor types, resulting in more accurate and coherent predictions.

Our model is empirically validated by means of the complex structured prediction task of image segmentation on the MSRC-21, KITTI, and SIFT Flow benchmarks. The results demonstrate that integrated inference and learning, and/or using neural factors, improves prediction accuracy over conventional SSVM training methods, such as *N*-slack cutting plane and subgradient descent optimization [1]. Furthermore, we demonstrate that our model is able to perform on par with current state-of-the-art segmentation models on the MSRC-21 benchmark.

* Corresponding author.
E-mail addresses: rein.houthooft@ugent.be (R. Houthooft), filip.deturck@ugent.be (F. De Turck).

http://dx.doi.org/10.1016/j.patcog.2016.03.014
0031-3203/© 2016 Elsevier Ltd. All rights reserved.

2. Related work

Although the combination of neural networks and structured or probabilistic graphical models dates back to the early 1990s [3,4], interest in this topic is resurging. Several recent works introduce nonlinear unary factors/potentials into structured models. For the task of image segmentation, Chen et al. [5] train a convolutional neural network as a unary classifier, followed by the training of a dense random field over the input pixels. Similarly, Farabet et al. [6] combine the output maps of a convolutional network with a CRF for image segmentation, while Li and Zemel [7] propose semisupervised maxmargin learning with nonlinear unary potentials. Contrary to these works, we trade the bifurcated training approach for integrated inference and training of unary and interaction factors. Several works [8–11] focus on linear-chain graphs, using an independently trained deep learning model whose output serves as unary input features. Contrary to these works, we focus on more general graphs. Other works suggest kernels towards nonlinear SSVMs [12,13]; we approach nonlinearity by representing SSVM factors by arbitrarily deep neural networks.

Do and Artières [14] propose a CRF in which potentials are represented by multilayer networks. The performance of their linear-chain probabilistic model is demonstrated by optical character and speech recognition using two-hidden-layer neural network outputs as unary potentials. Furthermore, joint inference and learning in linear-chain models is also proposed by Peng et al. [15], however, the application to more general graphs remains an open problem [16]. Contrary to these works, we propose a nonprobabilistic approach for general graphs by also modeling nonlinear interaction factors. More recently, Schwing and Urtasun [17] train a convolutional network as a unary classifier jointly with a fully connected CRF for the task of image segmentation, similar to [18,19]. Chen et al. [20] advocate a joint learning and reasoning approach, in which a structured model is probabilistically trained using loopy belief propagation for the task of optical character recognition and image tagging. Other related work includes Domke [21] who uses relaxations for combined message-passing and learning.

Other related work aiming to improve conventional SSVMs are the works of Wang et al. [22] and Lin et al. [23], in which a hierarchical part-based model is proposed for multiclass object recognition and shape detection, focusing on model reconfigurability through compositional alternatives in And-Or graphs. Liang et al. [24] propose the use of convolutional neural networks to model an end-to-end relation between input images and structured outputs in active template regression. Xu et al. [25] propose the learning of a structured model with multilayer deformable parts for action understanding, while Lu et al. [26] propose a hierarchical structured model for action segmentation.

Many of these works use probabilistic models that maximize the negative log-likelihood, such as [14,15]. In contrast, this paper takes a nonprobabilistic approach, wherein an SSVM is optimized via subgradient descent. The algorithm is altered to back-propagate SSVM loss errors, based on the ground truth and a loss-augmented prediction into the factors. Moreover, all factors are nonlinear functions, allowing the learning of complex patterns that originate from interaction features.

3. Methodology

In this section, essential SSVM background is introduced, after which integrated inference and back-propagation is explained for nonlinear unary factors. Finally, this notion is generalized into an SSVM model using only neural factors which are optimized by an alteration of subgradient descent.

3.1. Background

Traditional classification models are based on a prediction function $f:\mathcal{X}\rightarrow\mathbb{R}$ that outputs a scalar. In contrast, structured prediction models define a prediction function $f:\mathcal{X}\rightarrow\mathcal{Y}$, whose output can be arbitrarily structured. In this paper, this structure is represented by a vector in $\mathcal{Y}=\mathcal{L}^n$, with $\mathcal{L}\subset\mathbb{N}$ being a set of class labels. Structured models employ a compatibility function $g:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}$, parametrized by $w\in\mathbb{R}^D$. Prediction is done by solving the following maximization problem:

$$f(x)=\underset{y\in\mathcal{Y}}{\operatorname{argmax}}\, g(x,y;w). \tag{1}$$

This is called inference, i.e., obtaining the most-likely assignment of labels, which is similar to maximum-a-posteriori (MAP) inference in probabilistic models. Because of the combinatorial complexity of the output space $\mathcal{Y}$, the maximization problem in Eq. (1) is NP-hard [20]. Hence, it is important to impose on g some kind of regularity that can be exploited for inference. This can be done by ensuring that g corresponds to a nonprobabilistic factor graph, for which efficient inference techniques exist [1]. In general, g is linearly parametrized as a product of a weight vector w and a joint feature function $\varphi:\mathcal{X}\times\mathcal{Y}\rightarrow\mathbb{R}^D$.

Commonly, g decomposes as a sum of unary and interaction factors,[1] in which $\varphi=[(\varphi_U)^\top,(\varphi_I)^\top]^\top$. The functions φ_U and φ_I are then sums over all individual joint input–output features of the nodes $\psi_i(y,x)$ and interactions $\psi_{ij}(y,x)$ of the corresponding factor graph [1,12]. For example in the use case of Section 4, nodes are image regions, while interactions are connections between regions, each with their own joint feature vector. Data samples (x,y) are conform this graphical structure, i.e., x is composed of unary features x^U and interaction features x^I. Moreover, the unary and interaction parameters are generally concatenated as $w=[(w_U)^\top,(w_I)^\top]^\top$.

In this formulation, the unary features are defined as

$$\psi_i(y_i,x_i)=(\epsilon_i(x)^\top[y_i=m])^\top_{(m\in\mathcal{L})}, \tag{2}$$

while the interaction features for 2nd-order (edges) interactions are defined as

$$\psi_{ij}(y_i,y_j)=(\xi_{ij}(x)[y_i=m\wedge y_j=n])^\top_{((m,n)\in\mathcal{L}^2)}, \tag{3}$$

with $\epsilon_i(x)$ being the unary features corresponding to node i and $\xi_{ij}(x)$ the interaction features corresponding to interaction (edge) (i,j). Similarly, higher-order interaction features can be incorporated by extending this matrix into higher-order combinations of nodes, according to the interactions. In the experiments of this paper, unary features are bag-of-words features corresponding to each superpixel. Interaction features are also bag-of-words, but this time corresponding to all connected superpixels.

In an SSVM the compatibility function is linearly parametrized as $g(x,y;w)=\langle w,\varphi(x,y)\rangle$ and optimized effectively by minimizing an empirical estimate of the regularized structured risk

$$R(w)+\frac{\lambda}{N}\sum_{n=1}^{N}\Delta(y^n,f(x^n)), \tag{4}$$

with $\Delta:\mathcal{Y}\times\mathcal{Y}\rightarrow\mathbb{R}^+$ being a structured loss function for which holds $\forall y,y'\in\mathcal{Y}:\Delta(y,y')\geq 0$, $\Delta(y,y)=0$, and $\Delta(y',y)=\Delta(y,y')$; R a regularization function; λ the inverse of the regularization strength; for a set of N training samples $\{(x^n,y^n)\}_{n\in\{1,\ldots,N\}}\subset\mathcal{X}\times\mathcal{Y}$ that can be decomposed into V_n nodes and E_n interactions. In this paper, we make use of L_2-regularization, hence $R(w)=\frac{1}{2}\|w\|^2$. Furthermore, in line with our image segmentation use case in Section 4, the loss function is the class-weighted Hamming

[1] Maximizing g corresponds to minimizing the state of a nonprobabilistic factor graph, which factorizes into a product of factors. However, by operating in the log-domain, the state decomposes as a sum of factors.

Please cite this article as: R. Houthooft, F. De Turck, Integrated inference and learning of neural factors in structural support vector machines, Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2016.03.014