Distributed Non-Convex ADMM-inference in Large-scale Random Fields

Ondrej Miksik[1]
http://www.miksik.co.uk
Vibhav Vineet[1]
vibhav.vineet@gmail.com
Patrick Pérez[2]
patrick.perez@technicolor.com
Philip H. S. Torr[1]
http://www.robots.ox.ac.uk/~tvg/

[1] Department of Engineering Science
University of Oxford
Oxford, UK
[2] Technicolor Research & Innovation
Cesson Sévigné, FR

We propose a parallel and distributed algorithm for solving discrete labeling problems in large scale random fields. Our approach is motivated by the following observations: i) very large scale image and video processing problems, such as labeling dozens of million pixels with thousands of labels, are routinely faced in many application domains; ii) the computational complexity of the current state-of-the-art inference algorithms makes them impractical to solve such large scale problems; iii) modern parallel and distributed systems provide high computation power at low cost. At the core of our algorithm is a tree-based decomposition of the original optimization problem which is solved using a non convex form of the method of alternating direction method of multipliers (ADMM). This allows efficient parallel solving of resulting sub-problems. We evaluate the efficiency and accuracy offered by our algorithm on several benchmark low-level vision problems, on both CPU and Nvidia GPU. We consistently achieve a factor of speed-up compared to dual decomposition (DD) approach and other ADMM-based approaches.

Probabilistic graphical models such as the Markov Random Fields (MRF) and Conditional Random Fields (CRF), and related energy minimization based techniques have become ubiquitous in computer vision and image processing. They have been proven especially useful to solve a variety of important, high-dimensional, discrete inference problems. Examples include per-pixel object labelling, image denoising, image inpainting, disparity and optical flow estimation, etc. [2]. Their use nonetheless implies computational costs that are often not compatible with very large scale problems met today in many applications. This concern is at the heart of present work.

We first define a discrete random field $\mathbf{Y} = \{y_1, y_2, ..., y_N\}$ attached to the N nodes of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. Each random variable takes a label from a discrete space $\mathcal{L}$ of size L. We define $\mathcal{Y} = \mathcal{L}^N$ the set of all possible label assignments. This random field is a pairwise Markov Random Field (MRF) if there exists an energy function of the form

$$E(\mathbf{Y}) := \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j), \tag{1}$$

composed of unary and pairwise potentials. Finding the lowest cost labeling of the energy over $\mathcal{Y}$ is an NP-hard combinatorial problem which can be written as the Integer Linear Program (ILP)

$$\begin{aligned} \textbf{ILP}-\textbf{MRF}: \ \text{minimize} \quad & \sum_{i \in \mathcal{V}} \theta_i \cdot p_i + \sum_{(i,j) \in \mathcal{E}} \theta_{ij} \cdot q_{ij} \\ \text{with respect to} \quad & (p,q) \in \text{Marg}(\mathcal{G}). \end{aligned} \tag{2}$$

where θ_i, θ_{ij} are vectors of unary and pairwise potentials and p, q are corresponding binary indicators.

Following [1], we split the original graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ into S sub-graphs $\mathcal{G}_s = (\mathcal{V}_s, \mathcal{E}_s)$, $s = 1 \dots S$ and associate to each one auxiliary variables $p^s = \{p_i^s\}_{i \in \mathcal{V}_s}$ and $q^s = \{q_{ij}^s\}_{(i,j) \in \mathcal{E}_s}$, and potential parameters $\{\theta_i^s,\ i \in \mathcal{V}_s\}$ and $\{\theta_{ij}^s,\ (i,j) \in \mathcal{E}_s\}$, such that:

$$\sum_{s: i \in \mathcal{V}_s} \theta_i^s = \theta_i,\ \forall i \in \mathcal{V}; \qquad \sum_{s:(i,j) \in \mathcal{E}_s} \theta_{ij}^s = \theta_{ij},\ \forall (i,j) \in \mathcal{E}. \tag{3}$$

This implies that each node and each edge of the original graph must be covered by at least one sub-graph and that the sub-graphs can share freely nodes and edges and that the potentials on all shared vertices or edges of the sub-graphs sum to that of the original graph.

Given sub-graphs and associated parameters, we aim to replace the difficult inference problem (2) by a set of sub-problems that can be solved in parallel, while consistency between them is enforced in some way. Within the ADMM framework, there are several ways to achieve this goal. We choose to rely on "master" variables $p = \{p_i\}_{i \in \mathcal{V}}$ at the node level only. Thanks to constraints (3), it is easy to see that the original ILP-MRF problem can be written as

$$\begin{aligned} \textbf{DIP}-\textbf{MRF}: \ \text{minimize} \quad & \sum_{s=1}^{S} \left(\sum_{i \in \mathcal{V}_s} \theta_i^s \cdot p_i^s + \sum_{(i,j) \in \mathcal{E}_s} \theta_{ij}^s \cdot q_{ij}^s \right) \\ \text{with respect to} \quad & (p^s, q^s) \in \text{Marg}(\mathcal{G}_s),\ \forall s \\ & p_i \in \{0,1\}^L,\ \forall i \in \mathcal{V} \\ \text{subject to} \quad & p^s = p_{|s},\ \forall s \end{aligned} \tag{4}$$

where $p_{|s} = \{p_i\}_{i \in \mathcal{V}_s}$ denotes the sub-vector of p containing variables only for nodes of $s-$th sub-graph.

This problem can be turned into an unconstrained minimization problem by introducing the *augmented* Largrangian:

$$\begin{aligned} L_\rho(\{(p^s, q^s)\}, p, \{\lambda^s\}) = \sum_{s=1}^{S} \Big(& E_s(p^s, q^s; \theta^s) + \sum_{i \in \mathcal{V}_s} \lambda_i^s \cdot (p_i^s - p_i) + \\ & \frac{\rho}{2} \sum_{i \in \mathcal{V}_s} \|p_i^s - p_i\|_2^2 \Big) \end{aligned} \tag{5}$$

where $E_s(p^s, q^s; \theta^s) = \sum_{i \in \mathcal{V}_s} \theta_i^s \cdot p_i^s + \sum_{(i,j) \in \mathcal{E}_s} \theta_{ij}^s \cdot q_{ij}^s$, $p \in \{0,1\}^P$, $(p^s, q^s) \in \text{Marg}(\mathcal{G}_s)$ and $\lambda^s = \{\lambda_i^s\}_{i \in \mathcal{V}_s} \in R^{L \times |\mathcal{V}_s|}$. This is a consensus problem in that we essentially have multiple copies of the same variable that should take the value of the master.

Vector λ is the dual variable as in classic Lagrangian duality and ρ is a positive parameter. While the additional penalty destroys the separability as compared to classic Lagrangian, it helps solving dual problem efficiently. The ADMM approach conducts the joint optimization of augmented Lagrangian by alternating the following three steps:

$$(p^s, q^s)^{(t+1)} := \underset{(p^s, q^s) \in \text{Marg}(\mathcal{G}_s)}{\text{argmin}} \ L_\rho(\{(p^s, q^s)\}, p^{(t)}, \{\lambda^{s(t)}\}), \quad \forall s. \tag{6}$$

$$p_i^{(t+1)} := \mathcal{P}_{\text{Marg}(\mathcal{G})} \left(\frac{1}{|\mathcal{I}(i)|} \sum_{s \in \mathcal{I}(i)} \left(p_i^{s(t+1)} + \frac{1}{\rho} \lambda_i^{s(t)} \right) \right), \qquad \forall i \in \mathcal{V}, \tag{7}$$

$$\lambda_i^{s(t+1)} := \lambda_i^{s(t)} + \rho \left(p_i^{s(t+1)} - p_i^{(t+1)} \right), \quad \forall s,\ \forall i \in \mathcal{V}_s. \tag{8}$$

In this paper, we show how to solve such problem efficiently on a modern GPU. Our approach is easy to implement, since each sub-problem requires one call to a dynamic programming solver, and is highly suitable for modern GPUs with thousands of CUDA cores. Finally, we show empirically that our approach rapidly converges to a good quality estimates and is able to return a solution at any point in practice, which is important when developing interactive systems.

[1] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas. MRF energy minimization and beyond via dual decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(3):531–552, 2011.

[2] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappena, and C. Rother. A comparative study of energy minimization methods for markov random fields. *TPAMI*, 30 (6), 2007.