Conditional Random Fields as Recurrent Neural Networks for 3D Medical Imaging Segmentation

Miguel Monteiro[*]
INESC-ID &
Imperial College London
miguel.monteiro@imperial.ac.uk

Mário A. T. Figueiredo
Instituto de Telecomunicações &
Instituto Superior Técnico
mario.figueiredo@lx.it.pt

Arlindo L. Oliveira
INESC-ID &
Instituto Superior Técnico
aml@inesc-id.pt

Abstract

A novel way of describing Conditional Random Fields as Recurrent Neural Networks has been recently proposed that enables a Fully-Convolutional Neural Network with a Conditional Random Field on top to be jointly trained end-to-end with gradient descent methods. This algorithm is meant to improve the quality of semantic segmentation, however, the proposed implementation is only available for 2D RGB images. In this paper, we generalize the implementation to work with any number of spatial dimensions and reference channels. Furthermore, we test our implementation on two different 3D medical imaging datasets and observe that the performance differences were not statistically significant. We conclude that the performance increases observed in the 2D RGB case may not translate to these new domains and present possible explanations for this behaviour.

1 Introduction

Using a fully-connected Conditional Random Field (CRF) [1] after a Fully-Convolutional Neural Network (FCNN) [2] is one of the state-of-the art approaches in semantic segmentation [3]. The core idea behind this approach is that the FCNN will serve as a feature extractor that produces a coarse segmentation which is later refined by the CRF. Unlike a convolution layer which employs local filters, the CRF looks at every possible pair of pixels in the image, also known as a clique. The CRF is a graphical model where every clique is defined not only by the spatial distance between pixels but also by their distance in colour space. This allows the CRF to produce a segmentation with much sharper edges when compared to only using a FCNN. Recently, it was proposed a way of training the CRF and FCNN jointly by writing the CRF as a Recurrent Neural Network (RNN) which can be placed on top of FCNN, allowing the system to be trained end-to-end with gradient descent methods. We extend this approach to 3D medical images and make our implementation publicly available.

2 Methods

Consider an n-dimensional image with N hyper-voxels (pixels, voxels, etc…) on which we wish to perform semantic segmentation. We define X_j and I_j to be the label and colour value of hyper-voxel j, respectively. Consider a random field $\mathbf{X}$ defined over a set of variables $\{X_1, X_2, \ldots, X_N\}$ each taking a value from a set of labels $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$. Consider another random field $\mathbf{I}$ defined over the variables $\{I_1, I_2, \ldots, I_N\}$ where the domain of each variable is the possible colour values of a hyper-voxel in the image. A Conditional Random Field $(\mathbf{I}, \mathbf{X})$ is characterized by a Gibbs distribution $P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\left(-\sum_{c \in \mathcal{C}_\mathcal{G}} \phi_c(\mathbf{X}_c|\mathbf{I})\right)$, where $\mathcal{G}$ is a graph on $\mathbf{X}$ and each clique c in the set of cliques $\mathcal{C}_\mathcal{G}$ induces a potential ϕ_c. The Gibbs energy of labelling $\mathbf{x} \in \mathcal{L}^N$ is

[*] This work was developed while the author was at INESC-ID.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

$E(\mathbf{x}|\mathbf{I}) = \sum_{c \in \mathcal{C}_\mathcal{G}} \phi_c(\mathbf{X}_c|\mathbf{I})$ and the maximum a posteriori (MAP) labelling of the random field is $\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{L}^N} P(\mathbf{X}|\mathbf{I})$. $Z(\mathbf{I})$ is a normalization constant that ensures $P(\mathbf{X}|\mathbf{I})$ is a valid probability distribution. For notational convenience the conditioning will be omitted from now on, we define $\psi_c(\mathbf{x}_c)$ to denote $\phi_c(\mathbf{x}_c|\mathbf{I})$.

The Gibbs energy of the fully-connected pairwise CRF is the set of all unary and pairwise potentials [1]:

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j), \tag{1}$$

where i and j range from 1 to N. The unary potential $\psi_u(x_i)$ is computed independently for each hyper-voxel by a classifier. The pairwise potentials are given by:

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w^{(m)} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j), \tag{2}$$

where $k^{(m)}$ is a Gaussian kernel applied to arbitrary feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$, $w^{(m)}$ is linear combination of trainable weights and μ is a compatibility function between labels.

The feature vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ can be constructed from any feature space regarding the image. However, in this setting, they are chosen to take into account positions p_i and p_j, and the colour values I_i and I_j of the hyper-voxels in the image:

$$k(\mathbf{f}_i, \mathbf{f}_j) = w^{(1)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}\right)}_{\text{appearance kernel}} + w^{(2)} \underbrace{\exp\left(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}\right)}_{\text{smoothness kernel}}. \tag{3}$$

The parameters θ_α, θ_β and θ_γ are hyper-parameters that control the importance of the hyper-voxel difference in a specific feature space. This choice of $k(\mathbf{f}_i, \mathbf{f}_j)$ includes both an appearance kernel, which penalizes different labels for hyper-voxels that are close in space and color value, and a smoothness kernel which penalizes different labels for hyper-voxels close only in space. The compatibility function, μ, is a k by k matrix learnt from the data. It has zeros along its diagonal and trainable weights elsewhere for the model to be able to penalize different pairs of labels differently.

Since the direct computation of $P(\mathbf{X})$ is intractable we use the mean field approximation to compute the distribution $Q(\mathbf{X})$ that minimizes the KL-divergence $\mathbf{D}(Q||P)$, where Q can be written as a product of independent marginals, $Q(\mathbf{X}) = \prod_i Q_i(X_i)$. Minimizing the KL-divergence yields the following iterative inference algorithm:

Algorithm 1: CRF mean field approximation.

```
$Q_i(x_i) = \frac{1}{Z_i} \exp\{-\phi_u(x_i)\}$;                                Initialize $Q$
while not converged do
    $\tilde{Q}_i^{(m)}(l) \leftarrow \sum_{i \neq j} k^{(m)}(\mathbf{f}_i, \mathbf{f}_j) Q_j(l)$ for all $m$;      Message passing
    $\hat{Q}_i(x_i) \leftarrow \sum_{l \in \mathcal{L}} \mu^{(m)}(x_i, l) \sum_m w^{(m)} \tilde{Q}_i^{(m)}(l)$;      Compatibility transform
    $Q_i(x_i) \leftarrow \exp\{-\psi(xi) - \hat{Q}_i(x_i)\}$;                       Local update
    normalize $\hat{Q}_i(x_i)$
end
```

The key insight of the CRF as RNN paper [4] is that this inference algorithm can be written as a sequence of steps which can propagate gradient backwards like a RNN. The authors called this new layer a CRF as RNN layer which can be placed on top of existing FCNN architectures to improve the quality of semantic segmentation with the advantage of being trainable end-to-end with gradient descent methods. With the exception of the message passing step, most of these steps can be easily implemented in any existing deep learning framework. For this choice of kernels, however, the message passing step from every X_i to X_j requires high-dimensional filtering. A brute force implementation would have a time complexity of $\mathcal{O}(N^2)$. Therefore, we use the permutohedral lattice to approximate high-dimensional filtering [5] in linear time complexity $\mathcal{O}(N)$.

The main contribution of this work is the generalized implementation of the aforementioned algorithm. Our system works with any number of spatial dimensions and reference image channels as opposed to only 2D RGB images. Unfortunately, the message passing step which involves high-dimensional filtering cannot be easily implemented using existing operations. The available implementation of the permutohedral lattice was designed for 2D RGB images and only used CPU kernels. We have re-implemented the permutohedral lattice so that the implementation: supports any number of spatial dimensions and reference image channels; contains not only a CPU C++ kernel but also as a C++/CUDA kernel for fast computation in GPU; includes a TensorFlow Op wrapper so that it can be easily used in Python and incorporated in the computational graph. Our code for the permutohedral lattice (both CPU and GPU) implemented as a TensorFlow operation is available at `https://github.com/MiguelMonteiro/permutohedral_lattice` and the code for the CRF as RNN algorithm is available at `https://github.com/MiguelMonteiro/CRFasRNNLayer`.

3 Results and Discussion

To test whether using the CRF as RNN layer on top of a FCNN improved the segmentation quality for 3D medical images, we conducted two experiments. We used the V-Net [6] as the underlying network architecture for segmentation.

The PROMISE 2012 dataset [7] is a set of 50 three-dimensional mono-modal Magnetic Resonance Imaging (MRI) prostate images and the respective expert binary segmentation of the prostate. We re-sampled the images to have isotropic resolution of $1 \times 1 \times 2$ millimetres. We used 5-fold cross-validation and measured the performance using the Dice Coefficient (DC). The results for this experiment were DC $= 0.780 \pm 0.110$ and DC $= 0.767 \pm 0.109$ respectively with and without the CRF as RNN layer.

The Multimodal Brain Tumor Segmentation Challenge 2015 (BraTS 2015) [8] training dataset for High-Grade Glioma (HGG) is composed of 220 multimodal MRI images of brain tumors. All of the images have the same size and resolution ($1 \times 1 \times 1$ millimetres), and have 4 different channels (T1, T1c, T2 and Flair). The expert segmentation has 5 distinct labels: background, oedema, enhancing tumour core, non-enhancing tumour core and necrotic tumour core. The performance metrics for this task are: the whole tumour DC (DC_{WT}), which includes everything except the background; and the core tumour DC (DC_{CT}), which only includes the enhancing, non-enhancing and necrotic cores. For this experiment, we split the data-set into training and holdout set (85%/15%). The results for this experiment were $\text{DC}_{\text{WT}} = 0.738 \pm 0.105$; $\text{DC}_{\text{CT}} = 0.482 \pm 0.257$ and $\text{DC}_{\text{WT}} = 0.735 \pm 0.105$; $\text{DC}_{\text{CT}} = 0.488 \pm 0.244$ respectively with and without the CRF as RNN layer.

Looking at the results for both experiments and performing paired t-tests, we can conclude that the performance difference between using or not the CRF as RNN layer is not statistically significant. Hence, we conclude that, in these cases, using the CRF as RNN layer on top of a FCNN does not improve the segmentation quality. The fact that this algorithm seemingly works for 2D RBG images [4] but not for 3D MRI medical images can be due to a number of factors. Natural images tend to have much higher contrast and much sharper edges than MRI images. The edges between objects in natural images tend to be much more well defined (*e.g.* A building against a blue sky) than in MRI images (*e.g.* the oedema in a brain MRI is a slightly different shade of grey than the healthy region surrounding it). Since MRI images have much less contrast and tend to have blurry edges, the object of interest often fuses with the background slowly and seamlessly. Trained radiologists can use their knowledge of human anatomy and pathology in conjunction with the observed image to infer where the object of interest starts and ends. In contrast, the CRF only has access to differences between hyper-voxels, and these differences are zero or close to zero in low contrast, blurry edge environments. This means that there is much more sensitivity to the parameters θ_α, θ_β and θ_γ. Setting these parameters becomes very difficult especially when taking into account inter-image variability observed during training and the large size of the model which makes cross-validation unfeasible. Furthermore, the regions of interest to be segmented in our experiments were “local”, both the prostate and brain tumours fit inside the receptive field of the FCNN. This may not be the case in natural images where we might, for example, want to segment multiple birds out of the sky. For this reason, it is possible that the FCNN is already able to capture all of the relevant spatial and colour relations in the image and hence leaves CRF no room to improve.

Acknowledgments

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) under projects PRECISE - Accelerating progress toward the new era of precision medicine (LISBOA-01-0145-FEDER-016394), and UID/CEC/50021/2013.

References

[1] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[3] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, April 2018.

[4] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1529–1537, Washington, DC, USA, 2015. IEEE Computer Society.

[5] Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *Computer Graphics Forum*, volume 29, pages 753–762. Wiley Online Library, 2010.

[6] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[7] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, Robin Strand, Filip Malmberg, Yangming Ou, Christos Davatzikos, Matthias Kirschner, Florian Jung, Jing Yuan, Wu Qiu, Qinquan Gao, Philip "Eddie" Edwards, Bianca Maan, Ferdinand van der Heijden, Soumya Ghose, Jhimli Mitra, Jason Dowling, Dean Barratt, Henkjan Huisman, and Anant Madabhushi. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical Image Analysis*, 18(2):359 – 373, 2014.

[8] Bjoern Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elisabeth Gerstner, Marc-Andre Weber, Tal Arbel, Brian Avants, Nicholas Ayache, Patricia Buendia, Louis Collins, Nicolas Cordier, Jason Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Cagatay Demiralp, Christopher Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan Iftekharuddin, Raj Jena, Nigel John, Ender Konukoglu, Danial Lashkari, Jose Antonio Mariz, Raphael Meier, Sergio Pereira, Doina Precup, S. J. Price, Tammy Riklin-Raviv, Syed Reza, Michael Ryan, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Silva, Nuno Sousa, Nagesh Subbanna, Gabor Szekely, Thomas Taylor, Owen Thomas, Nicholas Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, page 33, 2014.