Meta-Learning for Instance-Level Data Association

Ronald Clark
Dyson Robotics Lab
Imperial College London
United Kingdom, SW7
ronald.clark@imperial.ac.uk

John McCormac
Dyson Robotics Lab
Imperial College London
United Kingdom, SW7
j.mccormac@imperial.ac.uk

Stefan Leutenegger
Dyson Robotics Lab
Imperial College London
United Kingdom, SW7
s.leutenegger@imperial.ac.uk

Andrew J. Davison
Dyson Robotics Lab
Imperial College London
United Kingdom, SW7
a.davison@imperial.ac.uk

Abstract

In this work we introduce a meta-leaning model for segmenting objects in a class-agnostic manner, that allows us to match and track novel objects of interest across multiple frames of video without specific instance models being known a priori. The work is motivated by the powerful implications that such a model can have have on visual SLAM and other computer vision tasks; in essence, a segmentation model that can efficiently learn new concepts from a very limited amount of data would be able to bridge the gap between biological and machine perception by allowing a perception system to perform robust and accurate object-level data association over changing environments and large time-spans. In this paper, we propose a novel combination of a metric deep learned embedding and a meta-learned optimizer to perform instance-level segmentation that is able to generalize to new classes with only a single training example. The cost function of our meta-learner is designed to predict the parameter updates for the segmentation network so that it can be trained on the fly to adapt to new instance classes while requiring only a few optimization iterations. We train and test our model on the DAVIS video-object segmentation dataset.

1 Introduction

Semantic segmentation models traditionally require a large number of training samples to learn to recognize new object classes. Therefore, although there has been much interest in integrating semantic object-level information into tasks such as visual SLAM, these approaches have not been able to do so very effectively as they inhibit the system's ability to generalize to novel environments with a new set of objects. Recently, however, much progress has been made in terms of meta-learning models which in principle allows learning systems to increase in efficiency as more data becomes available. Meta-learning models allow us to learn a training algorithm that is much more data-efficient than standard training methods by automatically finding an optimal learning strategy tailored for each task; the knowledge it learns is general to the higher-level problem being solved - not to the individual task at hand.

In this paper we introduce a meta-learning model for instance-level segmentation of objects in video frames. Given a single labelled image of a specific object instance, the goal of our model is to segment the object in the following video frames. The initial segmentation might come from, for example, an

object in a map which has been re-projected into the image and now we want to track the object over the coming frames through tracking-by-segmentation (also known as video-object-segmentation).

2 Related Work

Fine-tuning Fine-tuning is perhaps the most popular and simplest method for achieving our goal. With fine-tuning, a pre-trained semantic network network is adapted to a new instance by training parts of the network on the new instance. As we have mentioned, fine-tuning as a method for instance-level segmentation is a very naive approach as training on a single example can lead to extreme over-fitting and, futhermore, standard optimizers are sensitive to a multitude of hyperparameters that cannot be tuned given given only a single training sample. Online fine-tuning also requires hundreds of iterations of gradient decent to Even so, fine-tuning or online optimization has shown to perform competitively for video object segmentation [1, 2]. In our work, we build on this and propose to meta-learn an optimizer for training the instance-level segmentation network. This approaches, in essence, should allow us to combat the overfitting problem and also allows us to create an optimizer for our segmentation task which is less sensitive to hyperparameters.

One-shot learning A number of one-shot learning methods have been proposed for image classification [3] as well as for semantic segmentation [4] that are able to generalize to new classes given only a few examples. These approaches train two networks in tandem - one segmentation network and one parameter-generating network. The parameter-generating network is used to condition the segmentation network on the current instance of interest. It works by predicting the weight parameters of the segmentation network on a per-instance basis. The segmentation network performs the actual segmentation of the instance. In some sense, these networks perform an extreme case of meta-learning in which the parameter-generating network is tasked with "optimizing" the segmentaion network's parameters in a single step. However, this approach to meta-learning is rather limited as it significantly restricts the information available for finding the weights of the segmentation network. Our approach thus has capitalizes on the advantages of both the one-shot learning and fine-tuning paradigms and can be seen as a hybrid of the two. Our network requires only a few iterations to updates the weights and thus is faster than the vanilla fine-tuning method, and like the fine-tuning approach is not limited to the accuracy of a single-shot prediction of the parameters.

3 Problem Definition

We formalize the the instance segmentation problem as follows, using the same notation as in [4]. For each instance that we want to detect and track in a video sequence, we are given a set of image-mask pairs, $S = \{(I_s^i, Y_s^i(l))\}_{i=1}^k$. The segmentation of the first frame, $Y_s^i \in L_{test}$, is obtained by initializing it from a number of seed-point, using a manual annotation or by reprojecting a prior model of an object in a map created by a SLAM system. We operate in open-world conditions which means that the type of object instance $l \in L_{test}$ is an open-set and no predefined number of instance classes exist in the world. The segmentation function that we learn $f(I_q, S; \theta_f)$ is tasked with predicting a segmentation mask $\hat{M}_q$ of the current instance l given only the current image I_q and the seeding image-segmentation pair which usually comprises the first frame in the video.

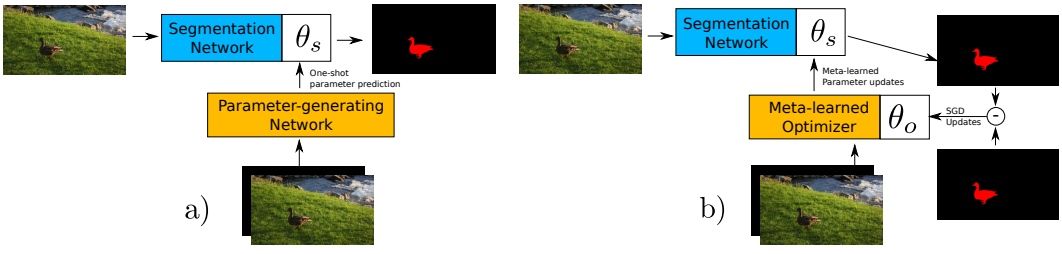

Figure 1: A comparison between (a) existing one-shot learning models for instance segmentation [4] and (b) our proposed meta-learning instance segmentation approach. Our model model learns to optimize the parameters in small number of iterations, and can use the single-shot prediction of the weights as an initialization.

Our model is based on learning an optimizer, $g(\theta_o)$ which finds the best $f(I_q, S; \theta_f)$ for an instance class by optimizing the parameters θ_f in a number of iterations. The algorithm is privy to a large set of instances L_{test} with labelled segmentation images across a sequence of frames during training. The test L_{test} and training L_{train} instance classes are not necessarily overlapping.

4 Proposed

Our meta-learned instance classifier operates on top of a robust deep-learned embedding of RGB images. Our embedding model is based on a standard fully-convolutional network (FCN) [5] with a pixel-wise contrastive loss as in [6, 7], while our meta-learning model for optimizing the parameters of the learner is based on the learning-to-learn by gradient descent method first introduced in [8].

4.1 Embedding Model

The fundamental task of the embedding model is to learn a feature space for the pixels so that the metric distances of instances of the same class lie close together while those pixels which correspond to different objects are far apart. Our embedding network, is initialized with weights that have been pre-trained on Imagenet and Pascal-VOC for calssification of semantic classes. As in [6, 7], our network takes as input an RGB image with 3 channels and outputs a d dimensional feature map representing the embedding, essentially converting each pixel into a high-dimensional feature vector.

We endow the feature space with a metric so that the closeness of two feature vecors can be computed. The metric needs to ensure that pixels which are of the same object has $d(a, b) \approx 0$, and that those which are of different instances or far apart have $d(a, b) \approx 1$. As in [6, 9] we use the following metric

$$d(a, b) = \frac{1}{1 + \exp(||f_a - f_b||_2^2)} \tag{1}$$

where f_a is the d dimensional embedding of a.The network is trained using a contrastive loss which has shown to perform very well for learning embeddings [9, 7]

$$\mathcal{L}_e = -\sum_{a,b \in K} \left[1_{\{y_p = y_q\}} \log(d(a, b)) \; + 1_{\{y_p \neq y_q\}} \log(1 - d(a, b)) \right] \tag{2}$$

Unlike standard losses used for training classification and segmentation networks the contrastive loss is defined between pairs of elements (in this case pixels). Evaluating this loss across the entire image is therefore prohibitively costly and infeasible for all but tiny images. Training is therefore carried out by sampling a small set of pixels K and evaluating the loss only at these points. The indicator function $1_{\{y_p \neq y_q\}}$ is set to 1 if the pixel is from the same class and 0 if it is from a different class. Class imbalance is accounted for by evenly sampling foreground and background classes.

4.2 Meta-learning Model

For meta-learning, we adopt the model first proposed in [8] and more recently popularised in [10]. In this model, the parameters of our segmentation network, θ_f are optimized using a standard update procedure described by

$$\theta_{f\,i+1} = \theta_{f\,i} + g_i \tag{3}$$

The per-iteration updates, g_i are predicted by a network which in this case is an LSTM-RNN

$$\begin{bmatrix} g_i \\ h_{i+1} \end{bmatrix} = \text{LSTM}_{cell}(\nabla_i, h_i, \theta_f; \theta_o) \tag{4}$$

where $\nabla_i = \frac{\partial \mathcal{L}_s}{\partial \theta_s}$ and LSTM_{cell} is a standard LSTM cell update function with hidden layer h_i.

In order to find the parameters of θ_o of the LSTM optimizer network, the following expectation loss is minimized

$$\mathcal{L}_{meta}(\theta_o) = \sum_{i=1}^{T} \sum_{k=1}^{K} \left[\mathcal{L}_s\big(f(I_k, S; \theta_{f\,i})\big) \right] \tag{5}$$

where the meta-loss is computed as the segmentation loss function, $\mathcal{L}_s$, summed over the T iterations of the learned optimization and the K frames of the video in which the instance is to be segmented. For each step i in the optimization we update the parameters of the segmentation model according to Eqn. 3. For the segmentation loss, $\mathcal{L}_s$, we simply use binary cross-entropy.

The meta-loss $\mathcal{L}_{meta}$ is optimized using standard gradient descent on the weights of the RNN optimizer. As our loss consists of variables which are updated recurrently over a number of timesteps, we use backpropogation through time to train the network. Backpropogation through time unrolls each step and updates the parameters by computing the gradients through the unrolled network. In our case we use RMSProp and use 50 iterations for our meta-optimization.

5 Training and Results

For convenience, we train our embedding network and segmentation network separately. To train our embedding, we make use of only synthetically rendered data in the form of the Scenenet RGB-D [11]. To train our meta-learned instance tracking-by-segmentation model, we make use of the densely-annotated video instance segmentation dataset (DAVIS 2017) [12, 13] which consists of 150 sequences with 10459 accurately segmented frames 349 different objects. For training we use only the training set and we perform testing on the publicly available validation set.

Measure		MSG [14]	NLC [15]	CUT [16]	FST [17]	MP-Net-F [18]	Memory [19]	Ours
$\mathcal{J}$	Mean	53.3	55.1	55.2	55.8	70.0	75.9	69.2
	Recall	61.6	55.8	57.5	64.9	85.0	89.1	82.0
$\mathcal{F}$	Mean	50.8	52.3	55.2	51.1	65.9	72.1	69.2
	Recall	60.0	51.9	61.0	51.6	79.2	83.4	75.3

Table 1: Comparison of existing methods [19] and our approach on DAVIS measured by intersection over union ($\mathcal{J}$) and F-measure ($\mathcal{F}$) .

The results of our approach on the DAVIS 2017 dataset are shown in Table 1 where $\mathcal{J}$ denotes region accuracy and $\mathcal{F}$ the contour accuracy. Our method performs slightly worse than the state-of-the-art approach presented in [19]. However, our method does not yet include any temporal information which can be integrated using pre-trained optical flow predictions and extending our segmentation model to an RNN. Our method does outperform other methods such as FST and MP-Net-F which have more complex architectures.

Figure 2: Qualitative evaluation on the DAVIS 2017 drift sequence. The instance segmentation model is trained using a single annotated frame from the start of the sequence. Annotations for subsequent frames are not shown to the model.

6 Conclusion

In this paper, we have introduced a meta-learning framework for performing instance-level segmentation in videos. Our model can generalize to novel instance types by seeing only one labelled training example for an instance. Our results are competitive with existing approaches which use more complex hand-crafted designs. For future work we inted to incorporate temporal constraints and will investigate the exciting possibilty of integrating these instance-level detections in a complete visual SLAM system. We believe that this dense tracking-by-segmentation model has many unexploited uses in both the computer vision and machine learning comunities.

References

[1] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR 2017*. IEEE, 2017.

[2] A. Newswanger and C. Xu. One-shot video object segmentation with iterative online fine-tuning. *The 2017 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2017.

[3] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016.

[4] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

[5] 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.

[6] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin P Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017.

[7] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2017.

[8] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001.

[9] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, pages 1857–1865, 2016.

[10] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pages 3981–3989, 2016.

[11] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. *arXiv preprint arXiv:1612.05079*, 2016.

[12] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017.

[13] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016.

[14] Thomas Brox and Jitendra Malik. Object segmentation by long term analysis of point trajectories. *Computer Vision–ECCV 2010*, pages 282–295, 2010.

[15] Alon Faktor and Michal Irani. Video segmentation by non-local consensus voting. In *BMVC*, volume 2, page 8, 2014.

[16] Margret Keuper, Siyu Tang, Yu Zhongjie, Bjoern Andres, Thomas Brox, and Bernt Schiele. A multi-cut formulation for joint segmentation and tracking of multiple objects. *arXiv preprint arXiv:1607.06317*, 2016.

[17] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784, 2013.

[18] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. *arXiv preprint arXiv:1612.07217*, 2016.

[19] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning video object segmentation with visual memory. *arXiv preprint arXiv:1704.05737*, 2017.